图灵数学 · 统计学丛书

统计机器学习界泰斗作品

Statistical Learning with Sparsity:
The Lasso and Generalizations

稀疏统计学习及其应用

[美] Trevor Hastie
Robert Tibshirani
Martin Wainwright 著

刘波 景鹏杰 译

全面介绍稀疏统计模型及其研究成果
用lasso模型解决大数据挖掘、机器学习等热点问题





中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

“作者们在书中研究分析了用一些统计模型中的稀疏特性来处理大数据的方法，主要关注lasso模型的求解算法和近年研究成果。”

——《数学文摘》

“本书涵盖了统计学的所有重要分支，每个主题都有基本问题的详尽介绍和求解算法，给出了基于稀疏性的分析方案。可以说，此书就是稀疏统计学习的标准教材。”

—— Anand Panangadan，加州大学富勒顿分校

“毋庸置疑，这本书是大数据技术领域重要著作。作为研究大数据的重要手段，lasso模型一直备受关注，但尚未有系统介绍其的相关资料，本书填补了这一空白，而且由领域内的三位大咖执笔，值得期待。”

—— Norm Matloff，加州大学戴维斯分校

Trevor Hastie 美国统计学家和计算机科学家，斯坦福大学统计学教授，英国皇家统计学会、国际数理统计协会和美国统计学会会士。Hastie参与开发了R中的大部分统计建模软件和环境，发明了主曲线和主曲面。

Robert Tibshirani 斯坦福大学统计学教授，国际数理统计协会、美国统计学会和加拿大皇家学会会士，1996年COPSS总统奖得主，提出lasso方法。Hastie和Tibshirani都是统计学习领域的泰山北斗，两人合著了*The Elements of Statistical Learning*，还合作讲授斯坦福大学的公开课“统计学习”。

Martin Wainwright 毕业于MIT，加州大学伯克利分校教授，以对统计与计算交叉学的理论和方法研究而闻名于学界，主要关注高维统计、机器学习、图模型和信息理论。2014年COPSS总统奖得主。



图灵社区：iTuring.cn
热线：(010) 51095186-600
分类建议 数学 / 统计学
人民邮电出版社网址：www.ptpress.com.cn







ISBN 978-7-115-47261-8
定价：89.00元



图灵数学·统计学丛书

Statistical Learning with Sparsity:
The Lasso and Generalizations

稀疏统计学习及其应用

[美] Trevor Hastie / Robert Tibshirani / Martin Wainwright 著

刘波 景鹏杰 译



人民邮电出版社
北京

图书在版编目(CIP)数据

稀疏统计学习及其应用/（美）特里瓦·哈斯蒂（Trevor Hastie），（美）罗伯特·蒂伯沙拉尼（Robert Tibshirani），（美）马丁·韦恩怀特（Martin Wainwright）著；刘波，景鹏杰 译 —北京：人民邮电出版社，2018.1
（图灵数学·统计学丛书）
ISBN 978-7-115-47261-8
Ⅰ.①稀… Ⅱ.①特… ②罗… ③马… ④刘… ⑤景… Ⅲ. ①统计学 Ⅳ.①C8
中国版本图书馆CIP数据核字（2017）第282216号

内 容 提 要

稀疏统计模型只具有少数非零参数或权重，经典地体现了化繁为简的理念，因而广泛应用于诸多领域。本书就稀疏性统计学习做出总结，以lasso方法为中心，层层推进，逐渐囊括其他方法，深入探讨诸多稀疏性问题的求解和应用；不仅包含大量的例子和清晰的图表，还附有文献注释和课后练习，是深入学习统计学知识的极佳参考。

本书适合算法、统计学和机器学习专业人士。


◆ 著　[美] Trevor Hastie, Robert Tibshirani, Martin Wainwright
译　刘 波 景鹏杰
责任编辑　朱 巍
责任印制　彭志环
◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编 100164　电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京隆昌伟业印刷有限公司印刷
◆ 开本：700×1000 1/16
印张：18.75　彩插：4
字数：368千字　2018年1月第1版
印数：1-3 000册　2018年1月北京第1次印刷
著作权合同登记号 图字：01-2015-8299号

定价：89.00元

读者服务热线：(010)51095186转600　印装质量热线：(010) 81055316

反盗版热线：(010)81055315

广告经营许可证：京东工商广登字20170147号


站在巨人的肩上
Standing on Shoulders of Giants
TURING
图灵教育
iTuring.cn

站在巨人的肩上
Standing on Shoulders of Giants

TURING
图灵教育

iTuring.cn

版权声明

Statistical Learning with Sparsity: The Lasso and Generalizations / by Trevor Hastie, Robert Tibshirani, Martin Wainwright / ISBN: 978-1-4987-1216-3.

谨以此书献给我们的父母亲：

Valerie Hastie、Patrick Hastie

Vera Tibshirani、Sami Tibshirani

Patricia Wainwright、John Wainwright

也献给我们的亲人：

Samantha、Timothy、Lynda

Charlie、Ryan、Jesss、Julie、Cheryl

Haruko、Hana

译　者　序

我们怀着无比敬仰的心情译完了这本书，因为本书的三位作者都是统计机器学习界的泰斗，他们在基于稀疏的统计学习理论方面的造诣有目共睹。能翻译此书是我们的荣幸。

随着大数据时代的到来，稀疏性成为研究大数据的重要手段。斯坦福大学的 Robert Tibshirani 于 1996 年首次提出将 ℓ_1 范数作为普通最小二乘的正则项，从而得到了著名的 lasso 模型。lasso 在拟和数据的同时，也利用 ℓ_1 范数让系数具有稀疏性来选择特征。这种稀疏性可增加模型的可解释性，并提高计算效率，这些特性对高维数据尤其重要。因此，ℓ_1 范数的这一特性很快就吸引了大批学者进行研究，并在各种机器学习模型中广泛使用，这些模型不仅仅包括广义线性模型，还包括图模型、信号处理中的压缩感知模型、矩阵填充模型等。ℓ_1 范数是一个凸函数，但不可微，因此，在求解以 ℓ_1 范数作为惩罚项的模型时通常比较麻烦。用来求解这类模型的常用算法是次梯度方法，但这种方法的效率通常不高，尤其在大规模的高维数据上更是如此。最近又出现了大量求解该问题的高效求解算法，比如坐标下降法等。

本书从最基本的 lasso 模型出发介绍了 lasso 模型的物理意义和相应的求解算法，然后介绍了 lasso 模型的推广和最新的研究成果。本书的特点是：(1) 采用深入浅出、图文并茂的方式介绍各种抽象理论；(2) 内容新颖，书中大量内容都是最近几年的研究成果，比如筛选规则等；(3) 涉及面广，书中的内容包含了与稀疏性相关的各个重要的研究领域，比如多元统计中的稀疏性问题等。因此，本书是广大机器学习研究人员、工程人员不可多得的参考书，也可用于研究生的机器学习课程的教材。

本书的第 1 章、第 2 章、第 5 章、第 7 章、第 8 章、第 10 章由重庆工商大学计算机科学与信息工程学院的刘波博士翻译，第 3 章、第 4 章、第 6 章、第 9 章、第 11 章由上海期货信息技术有限公司的景鹏杰翻译。

由于本书涉及的知识面广，内容也很新颖，因此许多术语尚无固定译法。我们虽然经过反复推敲和讨论，但仍然可能出现词不达意的情况。同时由于时间和能力有很，书中内容难免不出差错。若有问题或建议，读者可通过电子邮件 liubo7971@163.com 或 pjjing@foxmail.com 与我们联系，欢迎大家对本书的翻译进行指正或提出宝贵的建议。本书翻译的勘误信息会发布在 http://www.cnblogs.com/ml-cv/上，欢迎关注。

本书翻译过程得到如下项目资助:（1）重庆市教委研究项目“多核正则化机器学习理论研究”，项目号为 KJ130709;（2）重庆工商大学研究项目“基于多核学习的高维数据分析研究”，项目号为 2013-56-09;（3）大数据稀疏表示判别字典学习及其应用技术研究，项目号为 KJ1400612;（4）重庆工商大学研究生院教改项目“基于二维码的研究生互动教学改革”;（5）电子商务及供应链系统重庆市重点实验室项目“基于迹比率的特征选择及关键技术研究”，项目编号为 1456025。

感谢翻译过程中图灵公司的朱巍编辑给予的帮助和支持;感谢高敬雅老师，她给予了我们在统计学上的帮助;感谢重庆工商大学计信学院金融信息化专业的曾芳同学，她帮助我们录入了本书的公式。感谢刘波的妻子杨雪莉的支持，感谢刘波两个小女儿刘典、刘恩丫（此书完成时她才 7 个月大）对他的忍耐。

前　　言

在这本专著中，我们将概述基于稀疏性的统计学习的最新研究。稀疏的统计模型仅具有少数非零参数或权值。它代表了“少即是多”的经典情形：与稠密模型相比，稀疏模型更容易估计和解释。在这个大数据的时代，对一个人或目标进行度量的特征数量可能更多，而且有可能比观测样本数更多。借助稀疏性假设，我们能够解决这些问题，并从大数据集中提取有用的、可重复性模式。

这里所陈述的观点代表了统计学和机器学习方面全体研究人员的工作，我们感谢大家对这个令人激动的领域所作出的贡献。我们要特别感谢斯坦福大学和加州大学伯克利分校的同事，感谢我们的合作者，以及现在及过去在这个领域工作的同学们。他们分别是：Alekh Agarwal, Arash Amini, Francis Bach, Jacob Bien, Stephen Boyd, Andreas Buja, Emmanuel Candes, Alexandra Chouldechova, David Donoho, John Duchi, Brad Efron, Will Fithian, Jerome Friedman, Max G'Sell, Iain Johnstone, Michael Jordan, Ping Li, Po-Ling Loh, Michael Lim, Jason Lee, Richard Lockhart, Rahul Mazumder, Balasubramanian Narashimhan, Sahand Negahban, Guillaume Obozinski, Mee-Young Park, Junyang Qian, Garvesh Raskutti, Pradeep Ravikumar, Saharon Rosset, Prasad Santhanam, Noah Simon, Dennis Sun, Yukai Sun, Jonathan Taylor, Ryan Tibshirani①, Stefan Wager, Daniela Witten, Bin Yu, Yuchen Zhang, Ji Zhou, and Hui Zou。感谢我们的编辑 John Kimmel 对本书的建议和支持。

Trevor Hastie, Robert Tibshirani, 斯坦福大学
Martin Wainwright, 加州大学伯克利分校

①本书文献中提到的 Tibshirani_2 是 Ryan Tibshirani，而 Tibshirani 则是其父 Robert Tibshirani。

目　录

第1章 引　　言

“我从来都不记记分卡或击球率。我讨厌统计。我会把必须知道的东西记在脑海里。”

这段话是棒球投手 Dizzy Dean 说的，他曾在 1930 ~1947 年参加美国职业棒球大联盟的比赛。

一晃 75 年过去了，世界发生了很大的变化！如今，人们在科学、娱乐、商业和工业各领域收集和挖掘大量数据，并对其进行研究和应用。医学家们通过研究患者的基因组选择最佳的治疗方法，并由此了解这些疾病产生的根本原因。在线电影和网上书店会研究客户的评价，以便向他们推荐新的电影或书籍。社交网络会研究其会员及好友的资料，优化在线体验。而且，现在多数大联盟棒球队都有统计员收集和分析击球手和投手的详细信息，帮助球队经理和队员做出更好的决策。

由此可知，这个世界淹没在了数据中。而 Rutherford D. Roger 等人则说：

“我们淹没在了信息的海洋里，却渴求着知识。”

海量信息亟待整理，取其精华去其糟粕。为了成功完成这项工作，人们期望真实情况得以简化：也许人体内大约 30 000 个基因并非都与癌症的发展过程直接相关；也许只需要客户对 50 或 100 部电影做出评价就足以揭示他们的爱好；也许左撇子投手对付左撇子击球手会比较轻松。

这些情形背后都有简单性假设。**稀疏性**（sparsity）是简单性的一种形式，这也是本书的中心主题。简而言之，在一个稀疏统计模型中，仅有较少参数（也称预测子，predictor）在发挥重要作用。本书将介绍如何利用稀疏性来恢复一组数据中的基础信号。

最典型的例子是线性回归，即有 N 组观测值，每组观测值由一个输出变量 y_i 和 p 个相关预测子变量（也称特征）$x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ 所组成。线性回归的目标是通过预测子来预测输出值，既要正确预测将来的数据，又要找出哪些预测子在起重要作用。一个线性回归模型可设为：

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i \tag{1.1}$$

其中，β_0 和 $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ 是未知参数，e_i 为误差项。这些参数可用最小二乘法来估计，即最小化最小二乘目标函数：

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \tag{1.2}$$

通常，式 (1.2) 的所有最小二乘估计都不为零。若 p 很大，则最终模型会变得难以解释。事实上，若 $p > N$，最小二乘估计的结果并不唯一，有无穷多个解可使目标函数为零，而且大多数解都会过拟和（overfit）数据。

因此，这个估计过程需要进行约束（即**正则化**）。可采用 lasso（即 ℓ_1 **正则化**）回归，通过求解问题

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2, \quad \|\beta\|_1 \leqslant t \tag{1.3}$$

来估计参数，其中，$\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$ 是 β 的 ℓ_1 范数，t 是用户指定的参数。可将 t 看作参数向量的 ℓ_1 范数的预估值（budget），lasso 就在该预估值下寻找最好的拟和。

为什么要采用 ℓ_1 范数，而不采用 ℓ_2 范数或 ℓ_q 范数呢？这是因为 ℓ_1 范数很特别。如果预估值 t 足够小，lasso 会产生稀疏的解向量，即解向量仅有一些坐标不为零。若采用 ℓ_q 范数（其中 $q > 1$），则不会出现这种情况。对于 $q < 1$ 的情形，所得的解是稀疏的，但整个目标函数为非凸函数，求解该目标函数的计算量会很大。$q = 1$ 是产生凸问题的最小值，凸性极大地简化了计算，而且也满足稀疏性假设，这样就能处理具有上百万参数的问题。

因此，稀疏性的优势在于它可以解释拟和的模型，并且计算简单。除此以外，最近几年人们对该领域进行了深入的数学分析，发现稀疏性还有第三个优势，这个优势称为**押注稀疏性**（bet on sparsity）原理：

既然无法有效处理稠密问题， 倒不如在稀疏问题上寻找有效的处理方法。

具有稀疏性的统计学习

我们可从每个参数的信息量 N/p 来研究稀疏统计学习。如果 $p \gg N$ 且真实模型不稀疏，则样本数 N 太小，无法精确估计参数。若真实模型是稀疏的，也就是说真实模型仅含有 $k < N$ 个不为零的参数，则可使用本书所介绍的 lasso 和相关方法来有效估计这些参数。这可能有些让人惊讶，因为即使不知道 p 个参数向量的第 k 个元素是否为零，也可以这样做。当然，这样得到的结果相对而言不那么准确，而事实证明这样的准确性也相当不错了。

综上所述，对于数据分析师、计算机科学家和理论家而言，稀疏统计建模是令人兴奋的领域，并且也很实用。图 1-1 就展示了这样的例子。这些数据来自 349 个癌症病人样本中的 4718 个基因，通过量化这些基因表达的测量值而得到。这些癌症分为 15 类，包括膀胱癌、乳腺癌、中枢神经系统肿瘤，等等。这里的目标是通过

这 4718 个特征或部分特征来建立一个分类器，以预测癌症类别。所得到的分类器应该对独立样本（independent sample）有低的错误率，并且仅依赖于基因子集，以协助生物学基础研究。

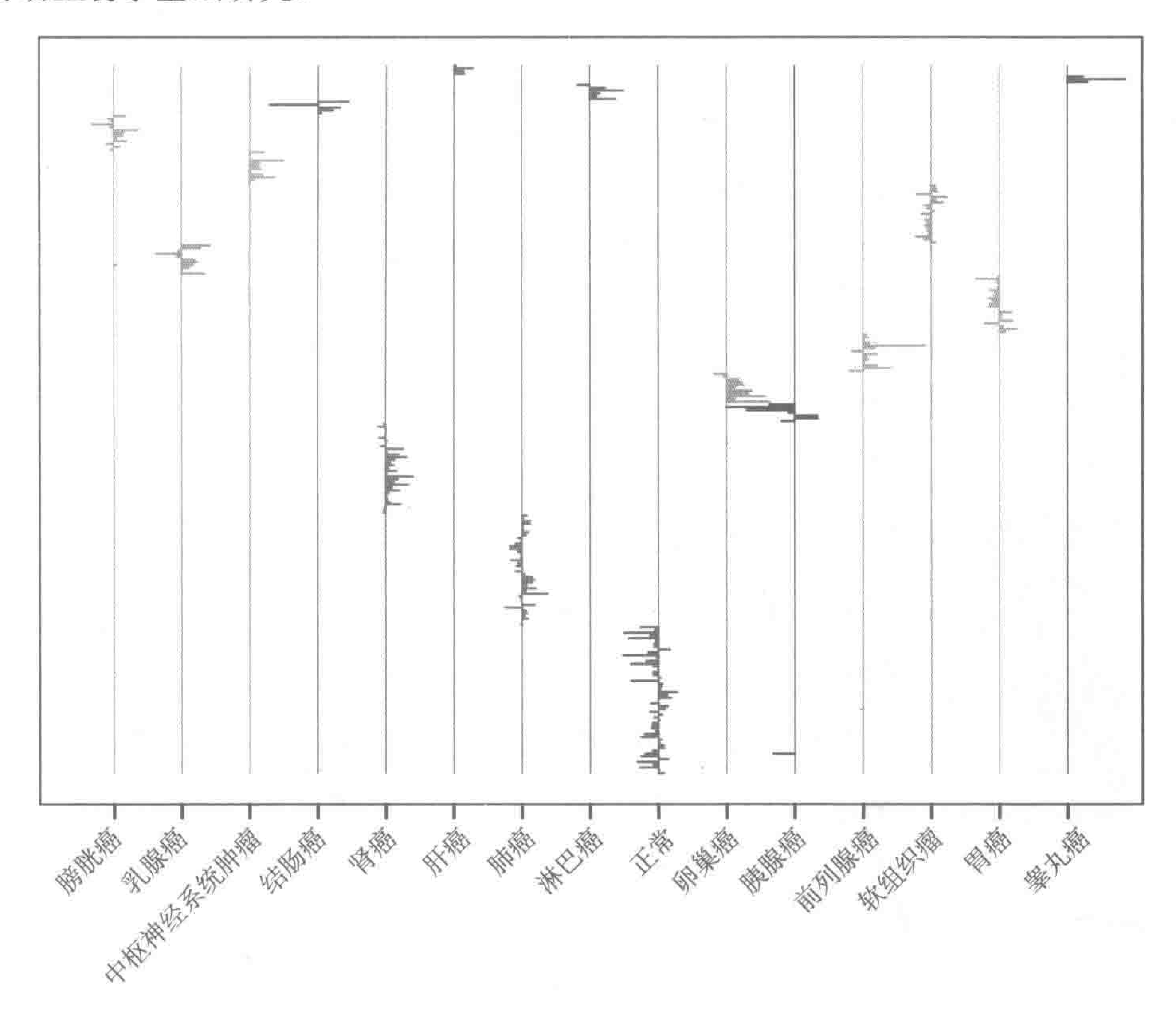


图 1-1 15 类癌症数据的基因表达（gene expression）：以一个 lasso 正则化的多项分类器（multinomial classifier）估计非零特征权重。这里显示了 4718 个基因中的 254 个，它们在 15 个类别中至少有一个非零权重。这些基因（未标记）按从上到下的顺序排列。指向右侧的线段表示正权重，指向左侧的线段表示负权重。从该图可看出，只需要极少的基因就可表示每个类

第 3 章会在这些数据上采用基于 lasso 正则化（lasso-regularized）的多项分类器来实现该目标。这会对 15 个类中的每一个生成 4718 个权重（或系数），以便在测试时进行区分。由于采用了 ℓ_1 惩罚，这些权重仅有一部分不为零（这取决于正则化参数的选取）。可通过交叉验证（cross-validated）来估计最优的正则化参数，图 1-1 显示了由此所得的权重。图中仅有 254 个基因有非零的权重。对该分类器进行的交叉验证，所得误差率约为 10%。也就是说，它能正确预测 90%的样本类别。相比之下，使用所有特征的标准支持向量机误差率（13%）稍高一些。lasso 所具有的

稀疏性会在不牺牲精度的情况下大幅减少特征数量。稀疏性也提高了计算效率：虽然可能要估计 $4178 \times 15 \approx 70\,000$ 个参数，但图 1-1 的整个计算在一个普通笔记本上不到一分钟就可完成。第 3 章和第 5 章所介绍的 glmnet 程序包可以完成相关计算。

图 1-2 展示了另一个例子（Candès and Wakin 2008），属于**压缩感知**（compressed sensing）领域。图 1-2a 是一幅具有上百万像素的图像。为了节省存储空间，图像可用小波基（wavelet basis）来表示，见图 1-2b。将最大的 25 000 个系数保留下来，其余的全部置为零，图 1-2c 是基于这些系数重构的图像，效果非常不错。这一切都归功于稀疏性：虽然图像看似复杂，但只有相对较少的小波基系数不为零。仅用 96 000 个不相关度量（incoherent measurement），原图像就可被完全恢复。压缩感知是一种强大的图像分析工具，详见第 10 章。

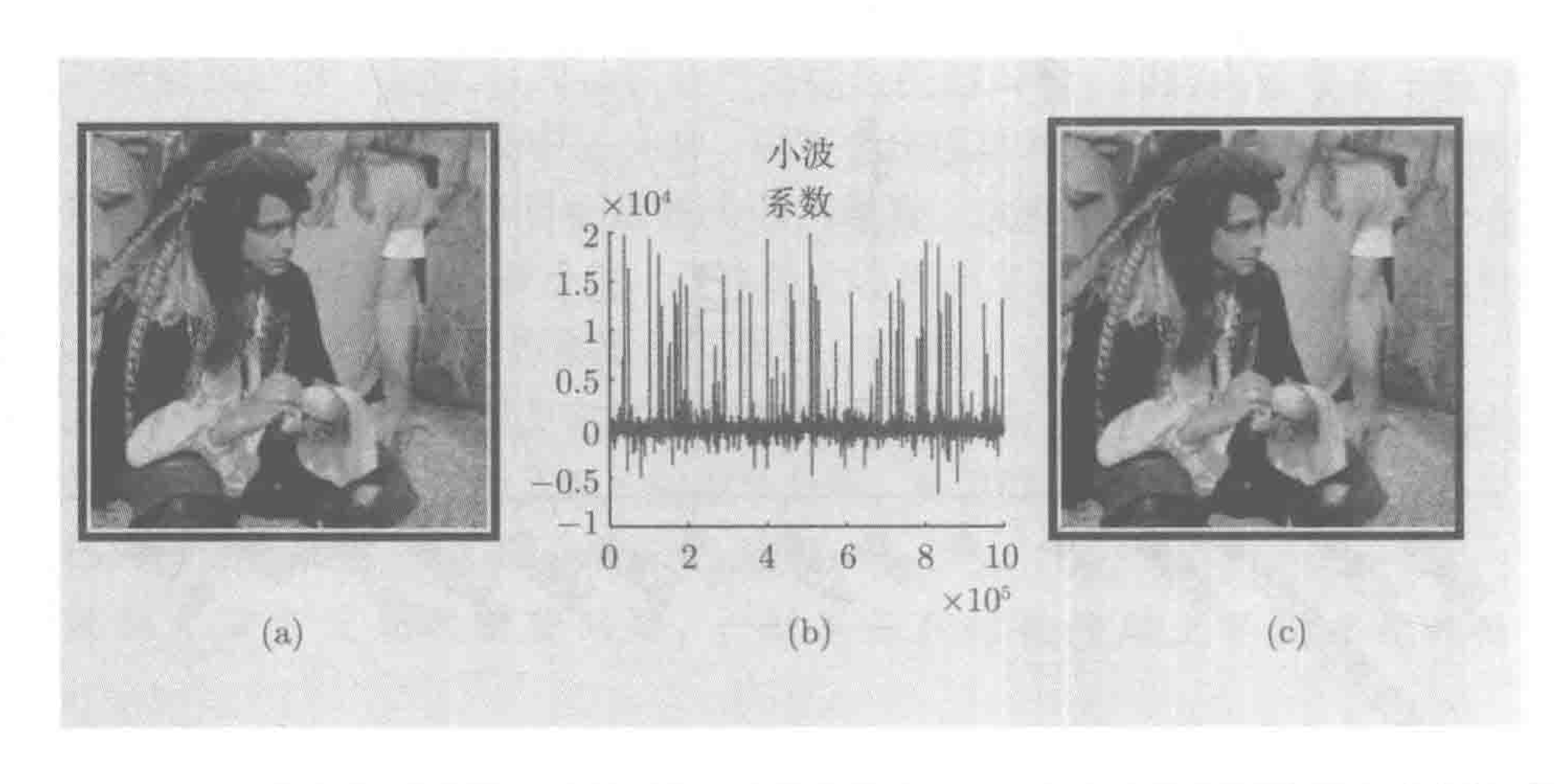


图 1-2　(a) 为具有上百万像素的原图像，其像素值为 [0,255]；(b) 为原图像的小波变换系数（随机排列这些系数，以加强其可视化效果）；少数小波系数得到了大部分信号能量；很多这样的图像都具有高压缩性；(c) 保留最大的 25 000 个小波系数（像素值的范围是 [0,255]），将其他系数置为零，由此重构原图像。原图像和重构图像之间几乎没有区别

本书意在总结稀疏统计模型最新的发展迅速的领域。第 2 章会介绍 lasso 线性回归，以及用来计算这种回归的简单坐标下降算法。第 3 章会介绍 ℓ_1 惩罚项在广义线性模型（比如多项式模型、生存模型以及支持向量机等）中的应用。第 4 章会介绍广义惩罚项，如弹性网（elastic net）和分组的 lasso。第 5 章会介绍优化问题的数值计算方法，重点介绍一阶方法，它能用于本书所讨论的大规模问题，第 6 章介绍拟和 lasso 模型的统计推断，包括 bootstrap、贝叶斯方法和一些最新的研究方法。第 7 章会介绍稀疏矩阵分解，并将这些方法应用到第 8 章的稀疏多元分析中。

第 9 章会介绍图和模型及其选择，而压缩感知会在第 10 章介绍。最后，第 11 章对 lasso 的理论成果进行概述。

需注意，本书会介绍监督学习问题和无监督学习问题。第 2 章、第 3 章、第 4 章和第 10 章讨论监督学习，而无监督学习会在第 7 章和第 8 进行讨论。

符号

为了清晰地讲解全书内容，这里对所有的数学符号进行统一规定。向量默认为列向量，因此 $\beta \in \mathbb{R}^p$ 为一个列向量，它的转置 β^{T} 为一个行向量。向量使用小写字母且不加粗，但 N 维向量加粗，其中 N 为样本大小。第 j 个变量的观测值是一个 N 维向量 $\boldsymbol{x}_j$，$\boldsymbol{y}$ 是一个 N 维响应向量。所有矩阵都会加粗，因此 $\boldsymbol{X}$ 表示 $N \times p$ 的观测矩阵。$\boldsymbol{\Theta}$ 为 $p \times p$ 的精度矩阵（precision matrix）。$x_i \in \mathbb{R}^p$ 表示第 i 个观测结果中的 p 个特征所构成的向量（即，x_i^{T} 为 $\boldsymbol{X}$ 的第 i 行），而 $\boldsymbol{x}_k$ 为 $\boldsymbol{X}$ 的第 k 列。

第 2 章　lasso 线性模型

本章将介绍线性回归的 lasso 估计，讲解其基本思想，以及一种简单的实现方法。lasso 估计与岭回归相关，也可看成一种贝叶斯估计。

2.1　引　　言

在线性回归中，样本 N 可表示为 $\{(x_i, y_i)\}_{i=1}^N$，其中，$x_i = (x_{i1}, \ldots, x_{ip})$ 是一个 p 维特征向量（也称为预测子向量），$y_i \in \mathbb{R}$ 是相应的响应变量。这里的目标是通过特征向量的线性组合

$$\eta(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \tag{2.1}$$

来估计响应变量 y_i。本模型的参数是回归权重向量 $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ 和截距（也称“偏置”项）$\beta_0 \in \mathbb{R}$。

对 (β_0, β) 的“最小二乘”估计通过最小化平方误差损失函数而得到：

$$\underset{\beta_0, \beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \tag{2.2}$$

替代最小二乘估计有两个主要原因。第一个原因是为了提高**预测精度**：最小二乘估计通常偏差小，但方差很大，有时可通过缩小回归系数范围或将一些系数置为 0 来提高预测精度。这种方法会让偏差变大，但同时会减少预测值的方差，因此能提高整体的预测精度（按均方误差度量）。第二个原因是为了满足**可解释性**：人们常常希望在大量特征中找出极具影响力的特征子集。

本章将主要讨论 lasso 方法，它将式（2.2）的最小二乘的损失函数与 ℓ_1 约束（系数绝对值之和）相结合。相对于最小二乘而言，这种约束可以收缩系数，甚至可将一些系数置零 [①]。lasso 为线性回归提供了一种自动选择模型的方法。此外，与

---

① lasso 是指一端有绳套的长绳，用来捕捉马和牛。这里是一个比喻，该方法会“套住”模型的系数。最初的 lasso 论文是 Tibshirani (1996)，正式命名为“lasso”是取“Least Absolute Selection and Shrinkage Operator ”首字母而成的。

发音：lasso 的美式发音往往被念成“lass-oh”（oh 就如山羊的叫声），而英式的 lasso 发音则为“lass-oo”。牛津英语词典（第 2 版，1965）指出：“lasso is pronounced lăsoo by those who use it, and by most English people too”。

其他模型选择方法不同，该方法得到的优化问题是凸的，这样能有效求解大规模问题。

2.2 lasso 估 计

对于具有 N 个预测子-响应变量对的数据集 $\{(x_i, y_i)\}_{i=1}^N$，lasso 给出了下面优化问题的解 $(\hat{\beta}_0, \hat{\beta})$：

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2\right\},\quad \text{其约束为}\sum_{j=1}^{p}|\beta_j|\leqslant t \tag{2.3}$$

约束 $\sum_{j=1}^p|\beta_j|\leqslant t$ 可简写成 ℓ_1 范数约束 $\|\beta\|_1\leqslant t$。式 (2.3) 通常会采用矩阵向量的形式来表示。如果用 $\boldsymbol{y}=(y_1,...,y_N)$ 表示 N 维响应向量，$\boldsymbol{X}$ 为一个 $N\times p$ 矩阵，并且 $x_i\in\mathbb{R}^p$ 为矩阵的第 i 行，则对于式 (2.3) 的优化问题，可以重新表示为

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2N}\|\boldsymbol{y}-\beta_0\mathbf{1}-\boldsymbol{X}\beta\|_2^2\right\},\quad \text{其约束为}\|\beta\|_1\leqslant t \tag{2.4}$$

其中，$\mathbf{1}$ 表示一个每个元素都为 1 的 N 维向量，$\|\cdot\|_2$ 为向量的欧氏范数。界 t 是一个预先设定的值，用来限制参数的绝对值之和。由于收缩参数会使约束模型更严格，这个预先设定的值会对拟合数据的优劣产生影响。该预设值必须通过其他方法来（例如交叉验证）来指定，这一点会在本章后面讨论。

通常，首先需要对预测子 $\boldsymbol{X}$ 进行归一化，即将列向量中心化 $\left(\frac{1}{N}\sum_{i=1}^N x_{ij}=0\right)$，而方差则要单位化 $\left(\frac{1}{N}\sum_{i=1}^N x_{ij}^2=1\right)$。如果不进行归一化，则 lasso 得到的结果会受到具体测量单位（比如英尺或米等）的影响。另一方面，如果单位统一，则不再进行归一化。为方便起见，通常假设输出值 y_i 已经中心化，即 $\left(\frac{1}{N}\sum_{i=1}^N y_i=0\right)$。这些中心化条件会带来方便，因为这样可在 lasso 优化中省略截距 β_0。如果在中心化的数据上能够得到 lasso 的最优解 $\hat{\beta}$，则非中心化数据得到的最优解也为 $\hat{\beta}$ 而截距 $\hat{\beta}_0$ 可通过下面的公式来计算：

$$\hat{\beta}_0=\bar{y}-\sum_{j=1}^{p}\bar{x}_j\hat{\beta}_j$$

其中，$\bar{y}$ 和 $\{\bar{x}_j\}_1^p$ 表示原始样本均值 ①。因此，本章会忽略 lasso 的截距 β_0。

方便起鉴，可用拉格朗日形式改写 lasso 问题：

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\left\{\frac{1}{2N}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2+\lambda\|\beta\|_1\right\} \tag{2.5}$$

① 这通常只是对有平方误差损失函数的线性回归成立，其他情形（如基于 lasso 的逻辑斯蒂回归）则不成立。

由拉格朗日的对偶可知，在式（2.3）约束形式和式（2.5）拉格朗日形式之间存在着一一对应关系：对一定范围内的 t，若 $\|\beta\|_1 \leqslant t$ 起作用约束，则每一个 t 值都有一个对应的 λ 值让式（2.5）与式（2.3）同解。反而言之，式（2.5）的解 $\hat{\beta}_\lambda$ 也是带约束 $t = \left\|\hat{\beta}_\lambda\right\|_1$ 的问题之解。

在一些 lasso 的相关文献中，式（2.3）和式（2.5）中的 $1/2N$ 通常会用 1/2 或 1 代替。但对式（2.3）没有什么影响，但对式（2.5）中的正则参数 λ 值有影响。经过归一化之后，不同大小的样本可以放在一起比较 λ 值（这对于交叉验证很有用）。

凸分析理论表明，若要保证满足式（2.5），则必然有

$$-\frac{1}{N}\langle \boldsymbol{x}_j, \boldsymbol{y} - \boldsymbol{X}\beta\rangle + \lambda \boldsymbol{s}_j = 0, \quad j = 1, \ldots, p \tag{2.6}$$

其中，若 $\beta_j \neq 0$，则每一个 s_j 都是一个未知量，等于 $\operatorname{sgn}(\beta_j)$，否则在 $[-1, 1]$ 区间取值，即 s_j 为绝对值函数的次梯度（第 5 章会详细介绍）。换句话说，式（2.5）的解 $\hat{\beta}$ 与式（2.6）的解 $(\hat{\beta}, \hat{\boldsymbol{s}})$ 相同。这个方程就是式（2.5）的 Karush-Kuhn-Tucker（KKT）条件。以次梯度来表示问题有助于设计求解算法。更多细节将在习题 2.3 和 2.4 中给出。

表 2-1 给出了另一个关于 lasso 的例子，其数据来自 Thomas（1990）。输出值为美国 50 个城市每百万居民的总犯罪率。这里有五个特征：警察年度经费（美元/居民）、25 岁及以上受过高中教育的居民比例、16~19 岁未入或辍学高中的居民比例、18~24 岁在校大学生的比例，以及 25 岁及以上至少受过四年大学教育的居民比例。这个示例只是为了阐释，但有助于展示 lasso 解的本质。通常情况下，lasso 对解决较大规模的问题（包括 $p \gg N$ 的“宽”数据问题）最为实用。

表 2-1　犯罪数据：美国 50 个城市犯罪率和 5 个特征

城市	经费	受过高中教育	没有受过高中教育	受过大学教育	大学毕业	犯罪率
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

图 2-1 左图表明，界 t 从左侧的 0 开始不断增加，达到右侧一个较大的值，这时 t 的取值不再对 lasso 算法产生影响。为了让最大化界为 1，需缩放图的横轴，相应的最小二乘估计为 $\tilde{\beta}$。由此可看出：在 t 的很多取值范围中，很多估计值为零，这就可将相应的特征从模型中删除。lasso 这样的模型选择特性是由 ℓ_1 约束（$\|\beta\|_1 \leqslant t$）背后的几何学原理决定的。右图的岭回归估计有助于理解这一点。岭回

归要比 lasso 出现得早，求解类似于式 (2.3) 的目标函数：

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2\right\},\quad 其约束为\sum_{j=1}^{p}\beta_j^2\leqslant t^2 \tag{2.7}$$

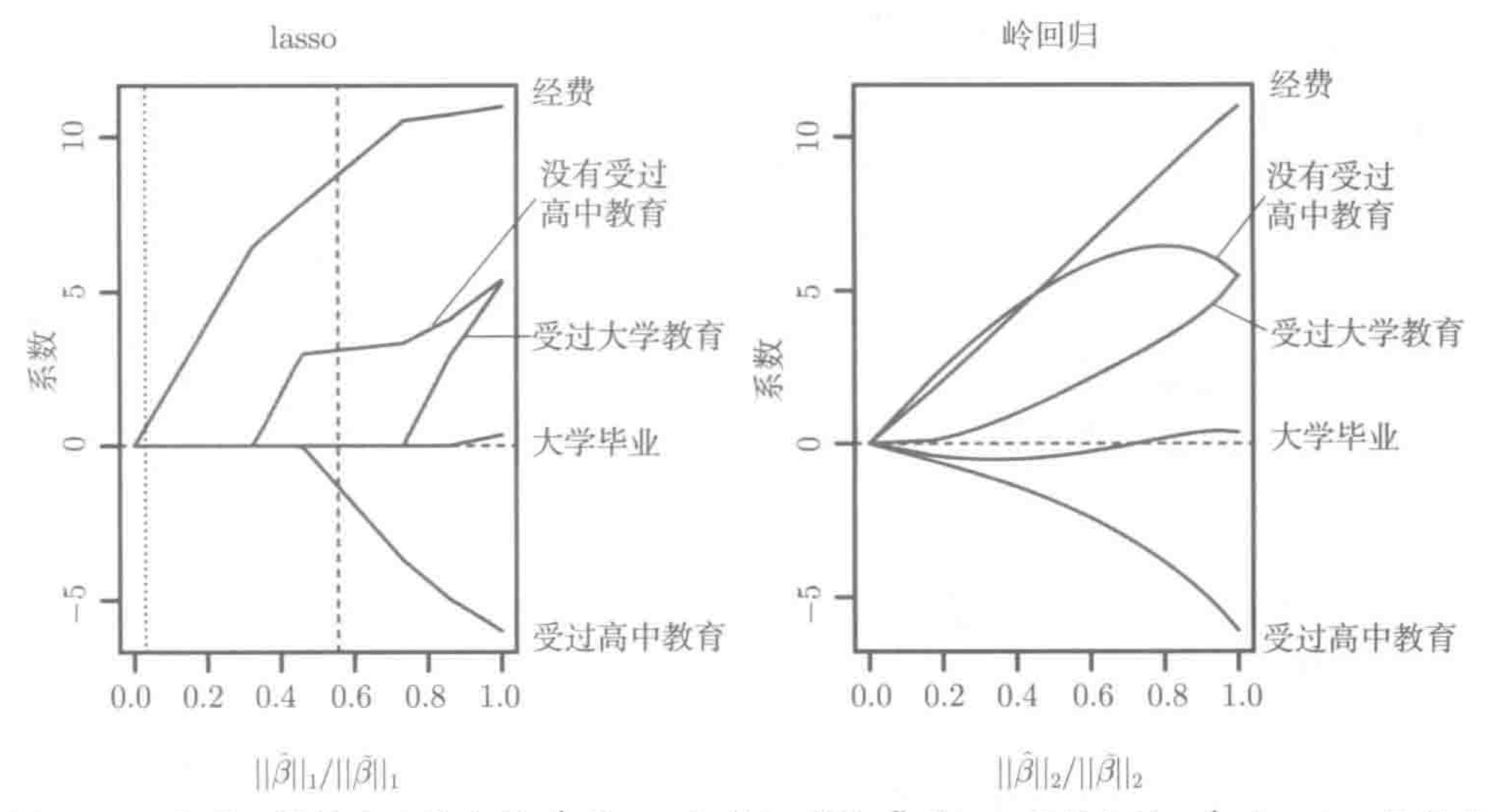


图 2-1 左图：横轴为系数向量 $\hat{\beta}$ 的 ℓ_1 范数与系数 $\tilde{\beta}$ 的 ℓ_1 范数之比。$\hat{\beta}$ 由 lasso 估计得出，$\tilde{\beta}$ 由无约束的最小二乘估计得出；右图：横轴为系数向量 $\hat{\beta}$ 的 ℓ_2 范数与系数 $\tilde{\beta}$ 的 ℓ_2 范数之比，$\hat{\beta}$ 由岭回归估计得到，$\tilde{\beta}$ 由无约束的最小二乘估计得到

岭回归的轮廓与 lasso 的轮廓大致相同，但岭回归只有左端的值为 0。图 2-2 对比了 lasso 和岭回归中所使用的约束。残差平方和函数拥有椭圆形轮廓，其中心为最小二乘估计的结果。岭回归的约束区域为圆盘 $\beta_1^2+\beta_2^2\leqslant t^2$，而 lasso 的约束区域为菱形 $|\beta_1|+|\beta_2|\leqslant t$。这两种方法找到的第一个点是椭圆形与约束区域相交的地方。与圆盘不同，菱形有拐角。如果解刚好在拐角处，那么就有一个参数 β_j 为 0。若 $p>2$，菱形就会变为偏菱形，并且有很多拐角、偏平边和面，所估计的参数有更多的机会为零（见图 4-2）。

我们称只有少量非零系数的模型为**稀疏模型**。ℓ_1 约束的一个重要性质是能得到稀疏解。这一点可以应用到许多统计模型中，同时也是本书的核心话题。

表 2-2 是对犯罪数据采用三种方法拟合的结果。lasso 中的 t 值通过交叉验证来确定，详见 2.3 节。表的左边是最小二乘拟合的结果，中间为 lasso 拟合的结果。右边的结果是这样得到的：用 lasso 结果非零系数对应的三个特征构成子集，然后在该子集上采用最小二乘法。最小二乘估计的标准误差可由公式计算得到，但 lasso 没有这样的公式，所以表 2-2 中间部分的标准误差只能利用 bootstrap 来获得（参

见习题 2.6，第 6 章会讨论后选择推断的实用新方法）。总体上，特征“经费”对犯罪率有着巨大影响，这意味着可以将更多警力资源集中在高犯罪地区；而其他的特征对犯罪率影响较小。

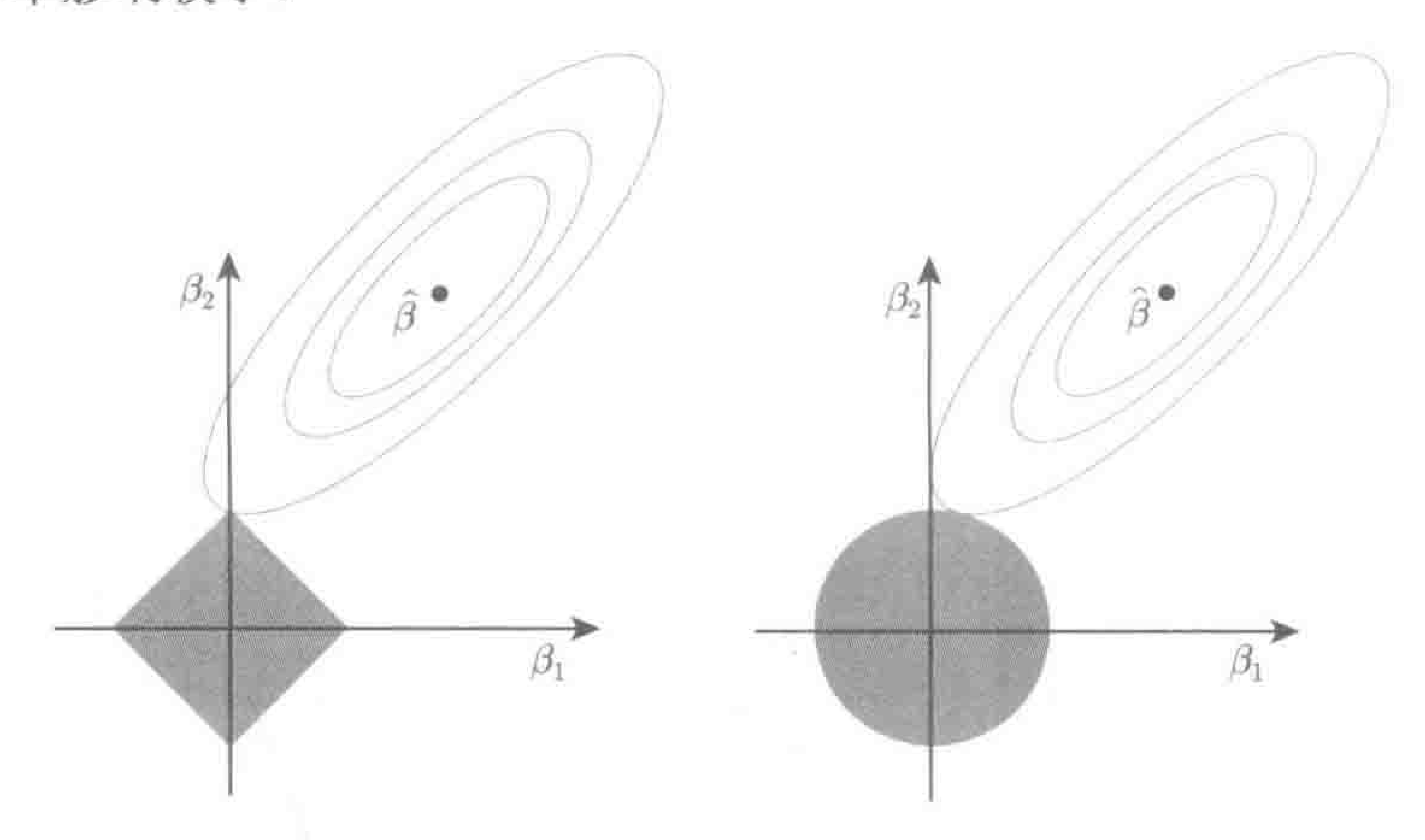


图 2-2　lasso（左图）和岭回归（右图）的估计图。实心区域分别是约束区域 $|\beta_1| + |\beta_2| \leqslant t$ 和 $\beta_1^2 + \beta_2^2 \leqslant t^2$。椭圆则是残差平方和函数的轮廓。点 $\hat{\beta}$ 是（非约束）最小二乘的估计值

表 2-2　犯罪数据分析结果。左边列出了最小二乘估计值、标准误差和二者的比率（Z-score）。中间及右边分别为 lasso 和在 lasso 得到的特征子集上进行最小二乘估计所得的结果

	LS coef	SE	Z	lasso	SE	Z	LS	SE	Z
经费	10.98	3.08	3.6	8.84	3.55	2.5	11.29	2.90	3.9
受过高中教育	−6.09	6.54	−0.9	−1.41	3.73	−0.4	−4.76	4.53	−1.1
没有受过高中教育	5.48	10.05	0.5	3.12	5.05	0.6	3.44	7.83	0.4
受过大学教育	0.38	4.42	0.1	0.0	–	–	0.0	–	–
大学毕业	5.50	13.75	0.4	0.0	–	–	0.0	–	–

注意，lasso 将五个系数中的两个都置为了零，而且相对于最小二乘估计，它会将这些非零系数向零收缩。另外，在三个特征构成的子集上的最小二乘会让系数远离零。lasso 的非零系数会向零偏移，所以表 2-2 右边这种去除偏移（debiasing）的方法往往可以改善模型的预测误差。这样的两阶段处理过程也称作松弛 lasso（relaxed lasso）（Meinshausen 2007）。

2.3　交叉验证和推断

在 lasso 目标函数式（2.3）中，界 t 可控制模型的复杂度。t 值越大，各个参数

的取值范围就越大，模型从而更有能力拟和训练数据。反之，t 值越小，参数的取值范围越小，模型从而更加稀疏，有更好的可解释性。假设暂时不考虑可解释性，对同一问题的独立测试数据，为了得到最好的预测精度，需要选择恰当的 t。这样的预测精度可以用来衡量模型的**泛化**能力。t 太小会令 lasso 难以获得数据中的主要信息，而太大又会导致过拟合（overfitting）。对于后一种情况，模型在拟合训练数据中的有用信息时，也拟合了无用的噪声。这两种情况都会让测试数据的预测误差偏大。因此，为了在这两种极端情况找到平衡，需要选择合适的 t 值，并且这过程使得一些系数为零。

为了估计最佳 t 值，可人为将数据集随机划分为训练集和测试集，并且采用**交叉验证**来获得测试集上的表现。具体而言，可以先随机将整个数据集分成若干组，假设组数为 $K > 1$。一般 K 可以选择 5 或者 10，有时候也可以为 N。将其中一组作为测试集，并指定剩下 $K-1$ 组为训练集。然后在训练集上对具有不同 t 值的 lasso 进行训练，并用测试集来测试训练好的模型响应值，同时记录下不同 t 值的均方预测误差（mean-squared prediction error）。这个过程共重复 K 次，以便让每组数据均有机会作为测试数据，而其他 $K-1$ 组作为训练数据。对于一系列 t 值，可通过这种方式来获得 K 个不同的预测误差估计。对每个 t 值，将这 K 个预测误差估计平均，从而得到**交叉验证的误差曲线**。

图 2-3 给出了犯罪数据的交叉验证的误差曲线，其中 K=10。图中纵轴为估

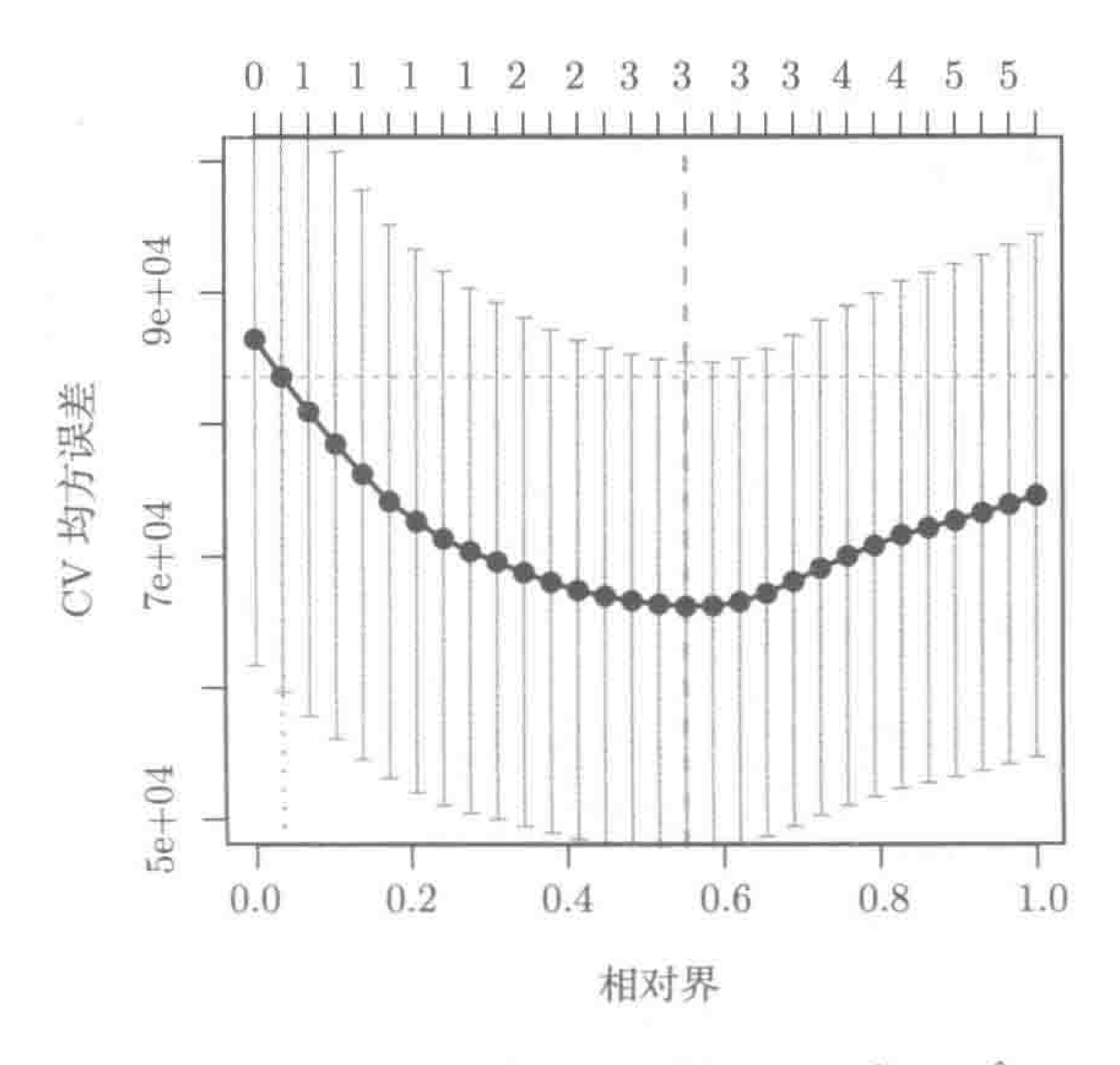


图 2-3　均方预测误差的交叉验证估计，它是相对的 ℓ_1 界 $\tilde{t} = \|\hat{\beta}(t)\|_1/\|\tilde{\beta}\|_1$ 的一个函数。这里的 $\hat{\beta}(t)$ 为约束取 t 时所得的 lasso 解，$\tilde{\beta}$ 则是普通最小二乘的解。这幅图显示了最小位置、逐点标准误差，以及“单位标准误差”的位置。由于样本数量 N 只有 50，标准误差会很大

计的均方预测误差，横轴为相对约束 $\tilde{t} = \| \hat{\beta}(t) \|_1 / \| \tilde{\beta} \|_1$。其中估计 $\hat{\beta}(t)$ 是约束取 t 值时所得的 lasso 解，$\tilde{\beta}$ 则是普通最小二乘的解。图 2-3 中的误差条是在交叉验证估计的预测误差上分别加减一个标准误差得出的。在均方预测误差的最小值（$\tilde{t}$=0.56）处有一条竖直的虚线，而另一条虚线位于“单位标准误差（$\tilde{t}$=0.03）”处。这是得到 CV 误差的最小 t 值，这个值所得到的交叉验证误差不会超过其最小值一个单位标准误差（one standard error）以上。图的最上面是每个模型的非零系数数量。从最上面这排数字可看出：有最小交叉验证误差的模型拥有三个特征，而单位标准误差模型只有一个特征。

注意，上面的交叉验证过程主要与参数 t 的界有关。人们可以对具有拉格朗日形式的式（2.5）进行参数 λ 的交叉验证。这两种方法会得到相似但不同的结果，因为在 t 与 λ 之间的映射与数据相关。

2.4　lasso 解的计算

lasso 问题属于凸规划，具体而言，它属于带有凸约束的二次规划（Quadratic Program，QP）。因此，有许多复杂的 QP 方法可以用来求解 lasso。但有一个特别简单且有效的计算方法，通过它可深入理解 lasso 的工作原理。为方便起见，可重写拉格朗日形式的目标函数：

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.8}$$

这里假定 y_i 和特征 x_{ij} 都进行过归一化处理，即：$\frac{1}{N}\sum_i y_i = 0, \frac{1}{N}\sum_i x_{ij} = 0, \frac{1}{N}\sum_i x_{ij}^2 = 1$，在这种情况下，截距 β_0 可以去掉。拉格朗日形式特别适合用简单的**坐标下降法**来求解目标函数。

2.4.1　基于单变量的软阈值法

下面首先考虑单个变量的情形，训练样本为 $\{(z_i, y_i)\}_{i=1}^N$（为方便起见，可将 z_i 看成 x_{ij}）。需要求解

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - z_i\beta)^2 + \lambda |\beta| \right\} \tag{2.9}$$

对于这种单变量的最小化问题，标准的求解方法是求关于 β 的梯度（一阶导数），并将其设置为零。但这样做还是有点复杂，因为绝对值函数 $|\beta|$ 在 $\beta = 0$ 处不可导。不过，可以直接处理式（2.9），从而得到

$$\hat{\beta}=\begin{cases}\frac{1}{N}\langle \boldsymbol{z},\boldsymbol{y}\rangle-\lambda, & \frac{1}{N}\langle \boldsymbol{z},\boldsymbol{y}\rangle>\lambda\\ 0, & \frac{1}{N}|\langle \boldsymbol{z},\boldsymbol{y}\rangle|\leqslant\lambda\\ \frac{1}{N}\langle \boldsymbol{z},\boldsymbol{y}\rangle+\lambda, & \frac{1}{N}\langle \boldsymbol{z},\boldsymbol{y}\rangle<-\lambda\end{cases} \tag{2.10}$$

这个结果的推导过程见习题 2.2，该结果可以写成

$$\hat{\beta}=\mathcal{S}_\lambda\left(\frac{1}{N}\langle \boldsymbol{z},\boldsymbol{y}\rangle\right) \tag{2.11}$$

这里的软阈值算子为

$$\mathcal{S}_\lambda(x)=\operatorname{sgn}(x)\,(|x|-\lambda)_+ \tag{2.12}$$

这个算子以 λ 使参数 x 向 0 平移。若 $|x|\leqslant\lambda$①，则得到 0（如图 2-4 所示）。注意，由于进行了归一化，即 $\frac{1}{N}\sum_i z_{ij}^2=1$，因而式（2.11）只是普通最小二乘估计 $\tilde{\beta}=\frac{1}{N}\langle \boldsymbol{z},\boldsymbol{y}\rangle$ 的软阈值版本。同样的结果可通过次梯度得到（见习题 2.3）。

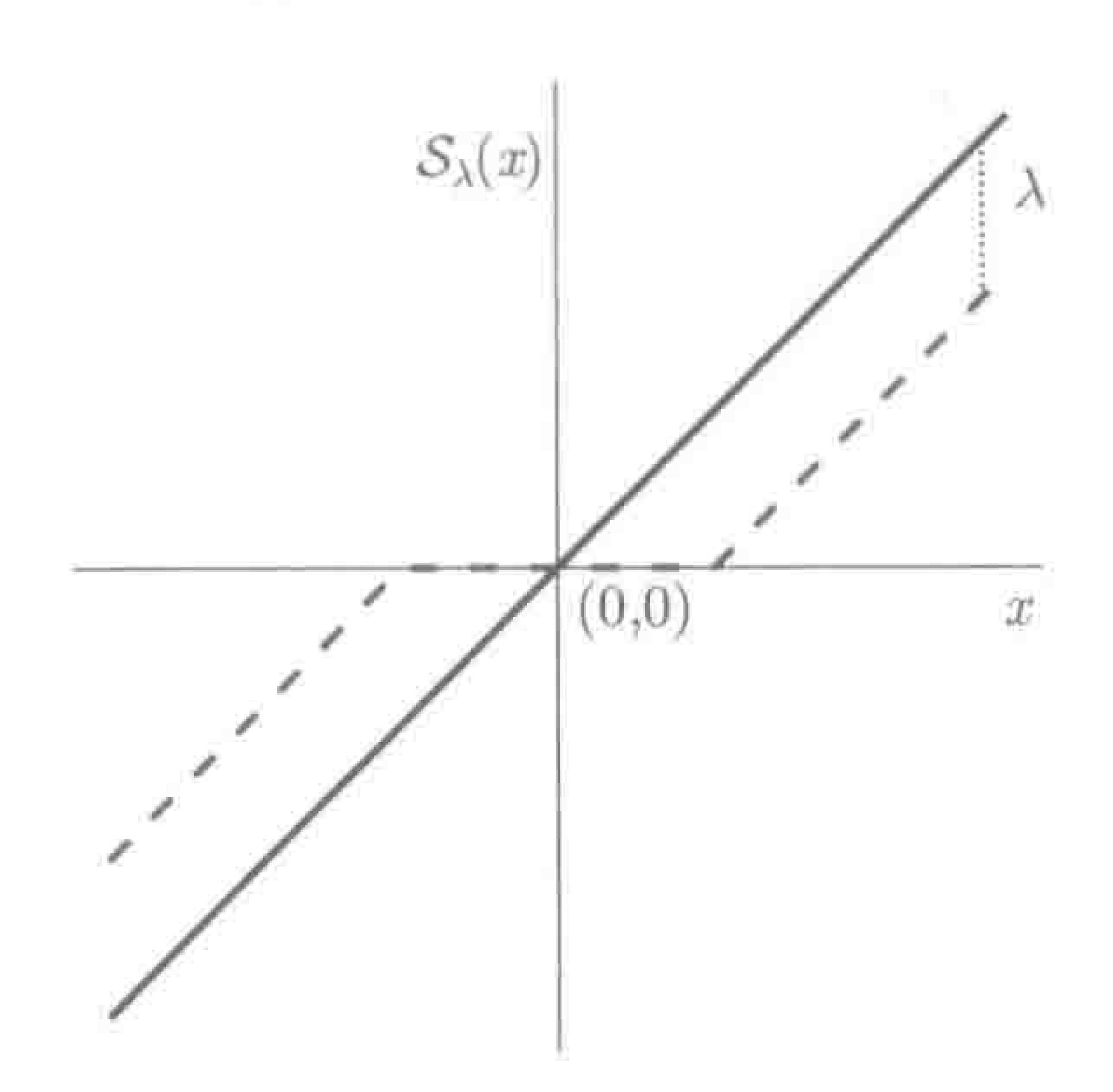


图 2-4 虚线为软阈值函数 $\mathcal{S}_\lambda(x)=\operatorname{sgn}(x)\,(|x|-\lambda)_+$，实线通过原点，斜率为 45°

2.4.2 基于多变量的循环坐标下降法

从单变量情形可看出：一般形式的 lasso 问题能够用一种简单的逐坐标（Coordinate wise）方法求解 [见式（2.5）]，也就是按某种固定顺序（比如：$j=1,2,\ldots,p$）依次求解。第 j 步令 $k\neq j$ 时的系数 $\hat{\beta}_k$ 不变，最小化 $k=j$ 时的目标函数，以更新系数 β_j。

因此，式（2.5）可重写为

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i-\sum_{k\neq j}x_{ik}\beta_k-x_{ij}\beta_j)^2+\lambda\sum_{k\neq j}|\beta_k|+\lambda\,|\beta_j| \tag{2.13}$$

① t_+ 表示 $t\in\mathbb{R}$ 且为正的部分，如果 $t>0$ 则为 t，否则为 0。

可以看到，每个 β_j 的解可以用偏残差（partial residual）表示，其定义为：$r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k$。在此，仅保留当前拟合结果的第 j 个变量。借助偏残差，第 j 个系数可以更新为

$$\hat{\beta}_j = \mathcal{S}_\lambda\left(\frac{1}{N}\left\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)}\right\rangle\right) \tag{2.14}$$

式 (2.15) 的等价形式为：

$$\hat{\beta} \leftarrow \mathcal{S}_\lambda\left(\hat{\beta}_j + \frac{1}{N}\left\langle \boldsymbol{x}_j, \boldsymbol{r}\right\rangle\right) \tag{2.15}$$

其中 $r_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$ 为整体残差（见习题 2.4）。整个算法会通过更新软阈值 [见式（2.14）] 的方式循环更新 $\hat{\beta}$ 的坐标以及由此得到的残差向量。

这个算法会得到最优解的原因在于，式（2.5）是 β 的凸函数，因此无局部极小值。该算法采用了**循环坐标下降法**，即每次沿一个坐标方向最小化凸目标函数。在相对宽松的条件下（一如本例），对凸函数采用这种逐坐标最小化方法即可收敛到全局最优解。要注意的是，这里有一定的条件限制，有些情形涉及不可分的惩罚函数，此时采用这种坐标下降方法就会失败。第 5 章会详细介绍这一点。

注意在式（2.5）中，若 $\lambda = 0$，则会得到普通最小二乘的解。从更新式（2.14）可以看出，该算法在每个特征上进行偏残差的单变量回归，直到最后收敛。当数据矩阵 $\boldsymbol{X}$ 满秩时，收敛的点为最小二乘的解。但这并不是效率很高的方法。

在实践中，人们通常感兴趣的不是在某个 λ 下得到的解，而是 λ 取值范围内所有解的路径（如图 2-1 所示）。一个合理的求解方法是以足够大的 λ 值开始，这个值下最优解为全零向量。如习题 2.1 所示，这需要取 $\lambda_{\max} = \max_j |\frac{1}{N}\langle \boldsymbol{x}_j, \boldsymbol{y}\rangle|$。然后略微缩小 λ 值，并采用坐标下降直至收敛为止；再次缩小 λ，并用上一次的解作为“热启动”（warm start），然后采用坐标下降法求解，直到收敛。通过这种方式，可对一系列 λ 的取值进行有效求解。这种方法称为**逐路径坐标下降**（pathwise coordinate descent）。

坐标下降法可以快速求解 lasso 问题，因为在每个坐标方向上的最小值可由式（2.14）显式得到，因此不需要沿每个坐标进行迭代搜索。另外，这样有利于得到问题的稀疏解：对于足够大的 λ，大多数系数将为零，且不会从零移走。5.4 节会针对活动集讨论大大提高算法效率的手段。

同伦方法（homotopy method）是另一类求解 lasso 问题的方法。这种方法从 0 开始，以连续方式得到解的整体路径。该路径实际上是分段线性的，如图 2-1 所示（可看成是 t 或 λ 的函数）。**最小角度回归**（Least Angle Regression，LARS）算法是一种同伦方法，能有效构建分段的线性路径，详见第 5 章。

2.4.3 软阈值与正交基

软阈值算子在 lasso 问题和信号去噪中至关重要。注意，若特征[1]间相互正交，即 $\frac{1}{N}\langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle = 0 (j \neq k)$，则上述坐标最小化方法会得到一种特别简单的形式。在这种情形下，更新后的式 (2.14) 可以大大简化，因为$\frac{1}{N}\langle \boldsymbol{x}_j, r^{(j)} \rangle = \frac{1}{N}\langle \boldsymbol{x}_j, \boldsymbol{y} \rangle$，$\hat{\beta}_j$ 就是基于单变量的最小二乘估计的软阈值化版本。这里说的"基于单变量的最小二乘估计"，即 $\boldsymbol{x}_j$ 中的每一个元素都是一个变量的样本值，$\boldsymbol{y}$ 则为 $\boldsymbol{x}_i$ 的输出。因此，在正交这样的特殊情形下，lasso 具有闭合解（close-form），即不需要迭代就能得到解。

小波是正交基的常见形式，用来对信号和图像进行平滑和压缩处理。在小波平滑中，有一组小波基表示数据，通过软阈值化小波系数来去噪。2.10 节和第 10 章会进一步讨论该问题。

2.5 自 由 度

假设有 p 个特征，所拟和的线性回归模型仅使用了 k 个特征。如果没有考虑响应变量就选出了这 k 个特征，则拟和过程会有 k 个自由度。对此，一种不严格的解释为：对于所有 k 个系数为零的检验假设，其标准检验统计量服从有 k 个自由度的卡方（Chi-squared）分布（假设误差方差 σ^2 已知）。

不过，如果这 k 个特征的选择关乎响应变量，例如：要在大小为 k 的所有子集上得到最小的训练误差，则需要拟和过程的自由度大于 k。我们称这种拟合过程具有自适应性，很明显 lasso 就是这样的算法。

与之类似，前向逐步（forward stepwise）过程会依次增加特征，从而最大程度缩小训练误差。在经过 k 步后，会得到自由度大于 k 的模型。由于这些原因，通常在拟合模型中，不能简单地将拟合模型中非零系数个数作为自由度值。但对于 lasso，可以将非零系数的数量作为自由度值。下面会介绍具体原因。

首先要准确定义自适应拟合模型中自由度的概念。假设有一个带误差的模型：

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, N \tag{2.16}$$

其中 f 是含有误差 ε_i 的未知函数，ε_i 独立同分布于 $(0, \sigma^2)$。$\hat{\boldsymbol{y}}$ 表示 N 个样本输出值，则可定义：

$$\mathrm{d}f(\hat{\boldsymbol{y}}) := \frac{1}{\sigma^2}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i) \tag{2.17}$$

这里的协方差用于衡量真实的输出变量 $\{y_i\}_{i=1}^N$ 与预测的输出变量之间的随机性。因此，自由度对应**自影响**（self-influence）的总数，自影响是对预测上每个输

[1] 此处的特征也称变量。—— 译者注

出的测量。模型越能拟合数据，自由度就越大。在一个固定的线性模型中，若所选择的 k 个特征与输出变量无关，则很容易证明 $\mathrm{d}f(\hat{\boldsymbol{y}}) = k$（见习题 2.7）。但对于自适应拟和而言，自由度通常会比 k 大。

对于有固定惩罚参数 λ 的 lasso，可以证明非零系数的数量 k_λ 是自由度[①]的无偏估计（Zou, Hastie and Tibshirani 2007 和 Tibshiran$_2$ and Taylor 2012），这有些不可思议。正如前面讨论得出的结论，特征选择方法（如前向逐步回归方法）在 k 步后得到的自由度会大于 k。由于前向逐步回归和 lasso 之间显然相似，因而 lasso 也具有这个简单的自由度性质。lasso 不仅选择特征（扩大自由度），而且相对于普通的最小二乘估计，它还会将这些特征对应的系数朝零收缩。这种收缩会让自由度下降到 k。这个结果非常有用，因为它对 lasso 路径中的任何一点都做了拟和量的定性度量。

在一般情形下，这种结果很难证明，在正交情形下，证明则比较容易。此时，lasso 估计可简化成单变量回归的软阈值版本。习题 2.8 会详细演示这一点，而且 6.3.1 节还会更进一步，基于 lasso 介绍检验特征显著性的**方差检验**。

2.6 lasso 解的唯一性

首先要注意，凸对偶理论可以证明：若 $\boldsymbol{X}$ 的列在一般位置（general position），则对于 $\lambda > 0$，lasso 问题（2.5）的解是唯一的。即使 $p \geqslant N$，这一点仍然成立，尽管任何 lasso 解的非零系数个数最多为 N（Rosset, Zhu and Hastie 2004，Tibshirani$_2$ 2013）。若数据矩阵 $\boldsymbol{X}$ 不是列满秩，最小二乘法拟合值唯一，但参数估计本身不唯一。非满秩情形在 $p \leqslant N$ 导致共线性时可能出现，在 $p > N$ 时则总会出现。对于后一种情形，$\hat{\beta}$ 的解有无限多个，这些解都会让训练误差为零。现在来考虑在 $\lambda > 0$ 的情形下，具有拉格朗日形式的 lasso 问题（2.5）。如习题 2.5 所示，拟合值 $\boldsymbol{X}\hat{\beta}$ 是唯一的，但所得的解 $\hat{\beta}$ 可能不唯一。比如有两个特征 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$，输出向量为 $\boldsymbol{y}$，假设在给定 λ 的情况下，lasso 的解为 $(\hat{\beta}_1, \hat{\beta}_2)$。如果有第三个特征 $\boldsymbol{x}_3 = \boldsymbol{x}_2$，则对于 $\alpha \in [0,1]$，向量 $\tilde{\beta}(\alpha) = (\hat{\beta}_1, \alpha\hat{\beta}_2, (1-\alpha)\hat{\beta}_2)$ 也是问题的解，且有 ℓ_1 范数 $\|\tilde{\beta}(\alpha)\|_1 = \|\hat{\beta}\|_1$。因此，这个模型（其中，$p \leqslant N$ 或 $p > N$）的解有无限多个。

在一般情况下，当 $\lambda > 0$ 时可证明，如果数据矩阵 $\boldsymbol{X}$ 的列是在**一般位置**，则 lasso 问题有唯一的解。准确地讲，列 $\{\boldsymbol{x}_j\}_{j=1}^p$ 在一般位置是指任何 $k < N$ 维的仿射子空间 $\mathbb{L} \subset \mathbb{R}^N$ 最多含有集合 $\{\pm\boldsymbol{x}_1, \pm\boldsymbol{x}_2, \ldots, \pm\boldsymbol{x}_p\}$ 的 $k+1$ 个元素，不包括正负点对（即这些点仅符号不同）。注意，在上一段的例子中，数据不在一般位置。如果 $\boldsymbol{X}$ 中的数据取自一个连续概率分布，则数据在一般位置的概率为 1，

① 对于 LAR 路径，会有一个更强的结论成立，若 $\boldsymbol{X}$ 满足某些条件，则在 k 步后，自由度会为 k。LAR 路径与 lasso 密切相关，这会在 5.6 节介绍。

因此 lasso 有唯一解。lasso 解的非唯一性仅出现在数据值离散的情形，比如分类特征的哑变量（dummy-value）编码时。这些结果以各种形式出现在文献中，文献 Tibshirani$_2$（2013）给出了这方面的总结。

注意，计算 lasso 解的数值算法通常在非唯一情形下会产生有效解。但这类算法得到的解与算法的具体实现细节有关。比如采用坐标下降时，初始值的选择会影响最终的解。

2.7 理论概述

lasso 求解算法有大量的理论，大多关注于 lasso 均方差（Mean-Squared-Error，MSE）的一致性和真实回归参数的非零支撑集（有时称为 sparsistency）求解。对于 MSE 的一致性，若 β^* 和 $\hat{\beta}$ 分别表示真实的参数值和由 lasso 估计的参数值，则当 $p, n \to \infty$ 时，会有很大的概率（Greenshtein and Ritov 2004, Bühlmann and van de Geer 2011，Chapter 6）得到如下结果：

$$\|\boldsymbol{X}(\hat{\beta} - \beta^*)\|_2^2/N \leqslant C \cdot \|\beta^*\|_1 \sqrt{\log(p)/N} \tag{2.18}$$

如果 $\|\beta^*\|_1 = o(\sqrt{N/\log(p)})$，则 lasso 对预测具有一致性。这意味着真正的参数向量必定相对于比率 $N/\log(p)$ 稀疏。其结果只需假定设计矩阵 $\boldsymbol{X}$ 固定。对非零支撑集的一致性求解需要在特征内和支撑集外的交叉相关（cross-correlation）上做更严格的假设。详见第 11 章。

2.8 非负 garrote

非负 garrote（Breiman 1995）[①]包括两个阶段，它与 lasso 关系紧密 [②]。给定初始的回归系数估计 $\tilde{\beta} \in \mathbb{R}^p$，则可求解下面的优化问题：

$$\underset{c \in \mathbb{R}^p}{\text{minimize}} \left\{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} c_j x_{ij} \tilde{\beta}_j)^2 \right\}, \quad \text{其约束为 } c \succeq 0 \text{ 且 } \|c\|_1 \leqslant t \tag{2.19}$$

其中，$c \succeq 0$ 表示向量的元素为非负。最后设 $\hat{\beta}_j = \hat{c}_j \cdot \tilde{\beta}_j, j = 1, \ldots, p$。这个目标函数有等价的拉格朗日形式，即将 $\lambda \|c\|_1$（$\lambda \geqslant 0$）作为正则惩罚项，然后加上非负约束。

在原论文（Breiman 1995）中，初始 $\tilde{\beta}$ 是通过普通最小二乘估计得到的。当 $p > N$ 时，估计不唯一。后来其他研究人员（Yuan and Lin 2006c，Zou 2006）证明：用其他初始估计（如 lasso、岭回归、弹性网等）时，非负 garrote 具有很好的性质。

---

① garrote 是一种用于通过绞杀或打破脖子执行死刑的设备。它是西班牙文，拼写成 garrotte 或 garotte 都可以。这里使用最早的论文 Breiman（1995）中的拼写。

② Breiman 的论文给 1996 年 Tibshirani 的论文提供了 lasso 灵感。

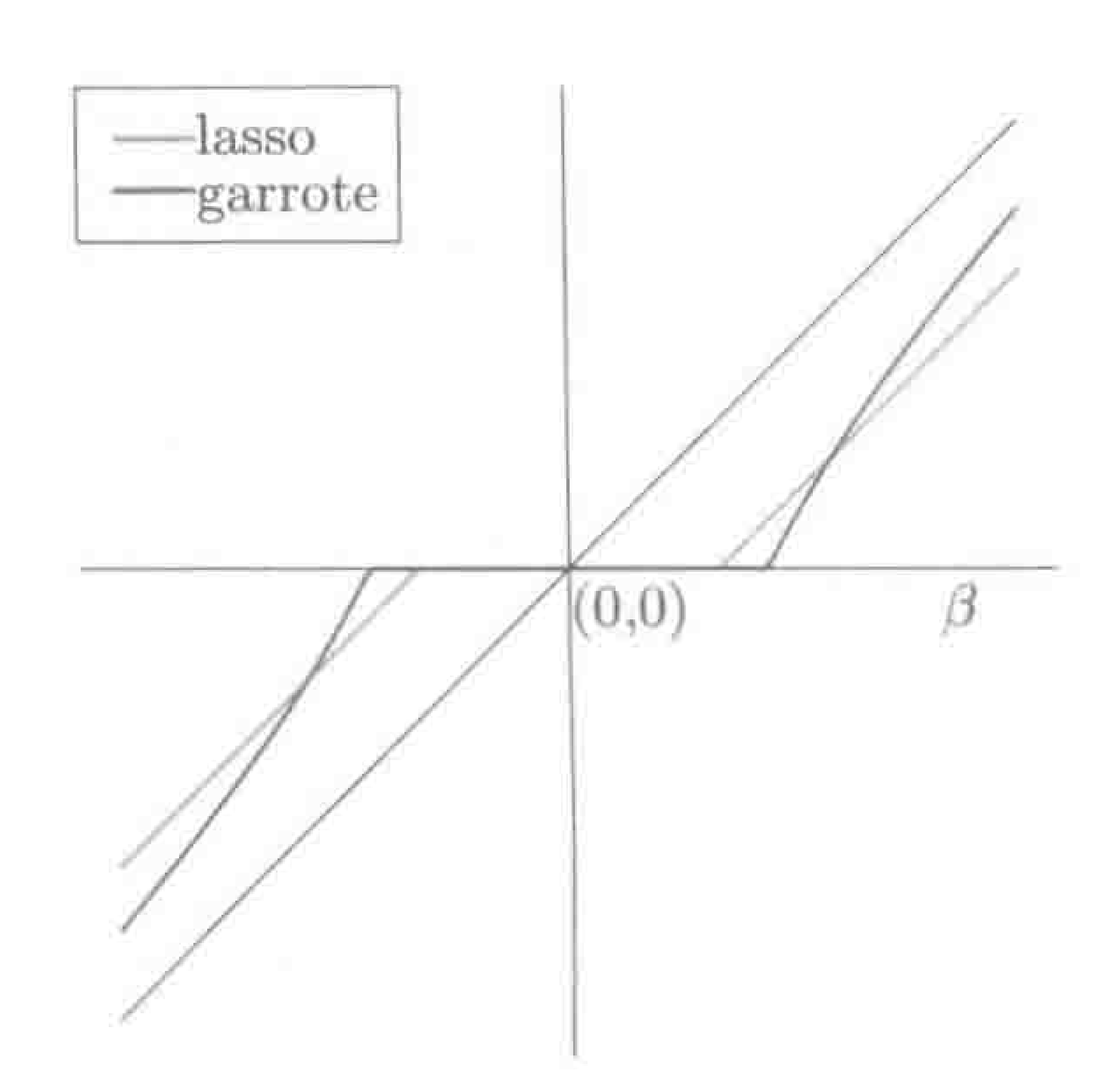


图 2-5 在单变量情形下，比较 lasso 和非负 garrote 的收缩行为。它们的 λ 大小不一样，lasso 中的 λ 为 2，garrote 的 λ 取 7。与 lasso 相比，garrote 会让 β 中较小的值收缩得更多，对于较大的值则相反

当 $\boldsymbol{X}$ 的列正交时，很容易得到非负 garrote 的解。假设 t 的取值范围可以满足等式 $\|c\|_1 = t$，则得到如下闭解：

$$\hat{c}_j = \left(1 - \frac{\lambda}{\tilde{\beta}_j^2}\right)_+, \quad j = 1, \ldots, p \tag{2.20}$$

其中，选取 λ 以满足 $\|\hat{c}\|_1 = t$。因此，如果系数 $\tilde{\beta}_j$ 较大，收缩因子会接近 1（即没有收缩），但如果它较小，则估计会朝零收缩。图 2-5 比较了 lasso 和非负 garrote 的收缩情况。后者展示了非凸惩罚的收缩情况（见 2.9 节和 4.6 节）。在非负 garrote 和**自适应 lasso**之间有着密切的联系，4.6 节和习题 4.26 还会讨论。

在此基础上，有文献（Yuan and Lin 2006c, Zhou 2006）证明：在比 lasso 要松的条件下，非负 garrote 是**路径一致**（path-consistent）的。如果初始估计是 $\sqrt{N}$ 一致的 [比如，基于最小二乘（$p < N$）、lasso、弹性网等的估计就具有这种一致性]，则路径一致性成立。“路径一致” 是指解的路径的某处包含着真实模型，解的路径由一组有序的 t 或 λ 决定。另一方面，非负 garrote 的参数估计的收敛性会比基于初始估计的要慢。

表 2-3 在数据矩阵 $\boldsymbol{X}$ 的列正交的情况下，式（2.22）中 β_j 的估计方法

q	估计方法	公式
0	最优子集	$\tilde{\beta}_j \cdot \mathrm{I}[\lvert\tilde{\beta}_j\rvert > \sqrt{2\lambda}]$
1	lasso	$\mathrm{sgn}(\tilde{\beta}_j)(\lvert\tilde{\beta}_j\rvert - \lambda)_+$
2	岭回归	$\tilde{\beta}_j / (1 + \lambda)$

2.9 ℓ_q 惩罚和贝叶斯估计

对于给定的实数 $q \geqslant 0$，有如下目标函数：

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|^q\right\} \tag{2.22}$$

上面这个式子，如果 $q=1$，则为 lasso；如果 $q=2$，则为岭回归。若 $q=0$，$\sum_{j=1}^{p}|\beta_j|^q$ 表示 β 中非零元素的个数，则求解式（2.22）就变成了获取最优子集。图 2-6 为两个特征（$p=2$）的情形下，这些惩罚项所对应的约束区域。式（2.21）中的 lasso 和岭回归情况为求解凸规划，所以很适合大规模问题。最优子集选择会得到非凸的组合优化问题，通常问题的特征数超过 $p=50$ 就不可解了。

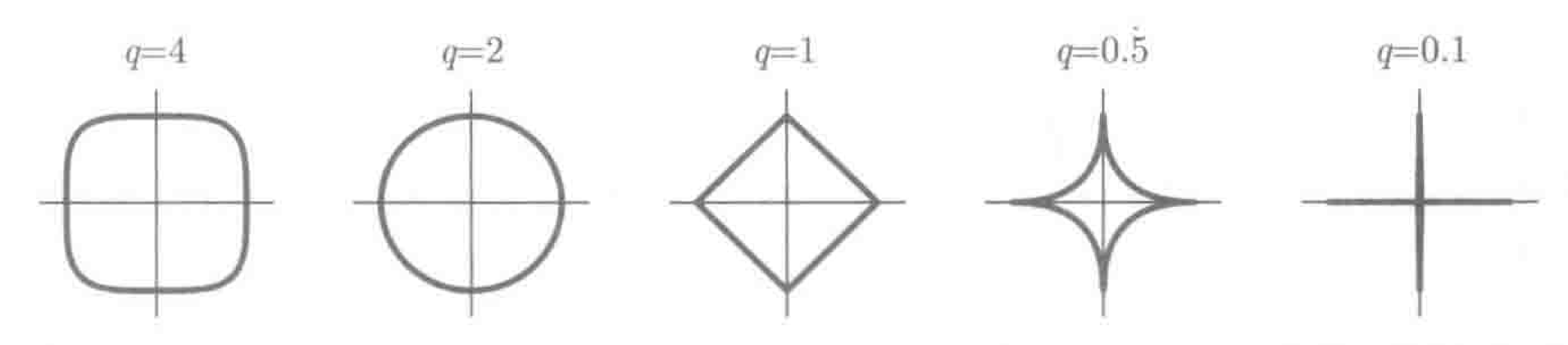


图 2-6　不同的 q 值所得到的约束区域 $\sum_{j=1}^{p}|\beta_j|^q\leqslant 1$，若 $q<1$，约束区域为非凸

若数据矩阵 $\boldsymbol{X}$ 的列正交，则这三个算法都有显式解。每种方法会对最小二乘估计 $\tilde{\beta}$（参见表 2-9）进行简单的逐坐标变换。岭回归进行了比例收缩。lasso 通过常数因子 λ 来对每个系数进行变换，并在零处截断，这也称为软阈值化。最优子集选择会采用硬阈值算子：保留大于 $\sqrt{2\lambda}$ 的系数，其他系数设为零。

lasso 很特殊，因为相应的 $q=1$，这是能得到凸约束区域的最小 q 值（最靠近最优子集）。在这个意义上，它是最接近最优子集选择的凸松弛问题。

这些估计方法也可以从贝叶斯的角度审视。可以认为 $|\beta_j|^q$ 与 β_j 的负对数先验密度成比例，图 2-6 所示的约束轮廓与参数先验分布的等轮廓（equi-contours）形状相同。注意，当 $q\leqslant 1$ 时，先验分布在坐标轴方向上更密集。若 $q=1$，则所有参数对应的先验分布是独立的双指数（或拉普拉斯）分布，其联合密度函数为 $(1/2\tau)\exp(-\|\beta\|_1/\tau)$（其中，$\tau=1/\lambda$）。因此，可以认为 lasso 估计是采用拉普拉斯分布作为先验分布的贝叶斯 MAP（Maximum Aposteriori，最大后验）估计；与后验分布均值不一样，它没有稀疏解。若从拉普拉斯先验分布所对应的后验分布采样，则不能得到稀疏向量。为了通过后验采样获得稀疏向量，需要从先验分布（该分布的概率大量集中在零周围）着手。6.1 节会介绍基于贝叶斯的 lasso 方法。

2.10 一些观点

lasso 使用了 ℓ_1 范数作为惩罚项，这种惩罚现已广泛应用于统计学、机器学习、工程、金融等多个领域。lasso 是由 Tibshirani（1996）受 Breiman（1995）的非负 garrote 启发而提出的。软阈值早些时候广泛用于基于小波的滤波中（Donoho and Johnstone 1994）。在信号处理中，小波通常用来替代傅里叶滤波，即同时考虑局部时频（local in time and frequency)。由于小波基是正交的，在 $\boldsymbol{X}$ 的列正交的情形下，小波滤波与 lasso 对应（见 2.4.1 节）。大约在 lasso 出现的同一时间，与小波密切相关的**基追踪**（basis pursuit）方法被提出（Chen, Donoho and Saunders 1998），扩展了小波滤波的思想，即在过完备（over-complete）基上，通过 ℓ_1 惩罚来对一个信号的稀疏表达进行搜索。这些都是正交框架的联合，因此不再完全相互正交。

从更广阔的角度审视，ℓ_1 正则化有其相当长的历史。例如，早在 1989 年，Donoho 和 Stark 就讨论了基于 ℓ_1 的恢复，并对不相关基（incoherent bases）提供了一些理论保证。他们甚至提到，在更早的 20 世纪 80 年代，地球科学界就有人进行了相关的研究（Oldenburg, Scheuer and Levy 1983，Santosa and Symes 1986）。在信号处理领域，Alliney 和 Ruzinsky 于 1994 年发表了与 ℓ_1 正则化相关的算法研究。还有其他研究人员也提出了相似的观点（Fuchs 2000）。Rish 和 Grabarnik（2014）则介绍了机器学习和信号处理中的稀疏方法。

在过去 10~15 年中，人们已经得到了 ℓ_1 惩罚的诸多良好性质，这些性质总结如下。

利于解释最终模型：ℓ_1 惩罚会使解变得稀疏且简单。

统计有效：在著作 *The Elements of Statistical Learning*（Hastie, Tibshirani and Friedman 2009）中，作者提出了押注稀疏性（bet-on-sparsity）原则，即假定真实信号是稀疏的，可通过 ℓ_1 惩罚来很好地求解。如果假定正确，那么我们就可以恢复真实信号。需要注意，稀疏性可保持在给定的基（特征集）或特征变换（比如一组小波基）中。但如果真实信号在所选择的基中不是稀疏的，则采用 ℓ_1 惩罚并不能得到好的效果。不过在这情况下，相对于贝叶斯误差，也没有更好的方法进行求解了。第 11 章会给出大量理论来支持这个观点。

计算高效：基于 ℓ_1 的惩罚项是凸的，假定稀疏性可得到显著的计算优势。如果有 100 个样本，每个样本有 100 万个特征，则必须估计 100 万个非零参数，这样计算量会非常大。但如果采用 lasso，则在解中最多有 100 个参数为非零，这会使计算更加容易。第 5 章会给出更多细节[①]。

本书剩余部分会介绍这一领域诸多令人兴奋的进展。

---

① 在 $p \gg N$ 时，岭回归也有类似的效率。

习　题

习题 2.1　若 lasso 估计的回归系数都为零，求证 λ 的最小值为：

$$\lambda_{\max} = \max_j \left| \frac{1}{N} \langle \boldsymbol{x_j}, \boldsymbol{y} \rangle \right|$$

习题 2.2　请以基于软阈值的算法（2.12）求解单变量 lasso 问题（2.9）。（注意：请不要利用次梯度）。

习题 2.3　**软阈值化和次梯度**。由于式（2.9）是一个凸函数，它的次梯度一定存在（详细介绍参见第 5 章），任何最优解必须满足次梯度方程：

$$-\frac{1}{N}\sum_{i=1}^{N}(y_i - z_i\beta)z_i + \lambda s = 0 \tag{2.22}$$

其中 s 是 $|\beta|$ 的次梯度。对于绝对值函数，其次梯度的形式为 $s \in \operatorname{sgn}(\beta)$，即当 $\beta \neq 0$ 时，$s = \operatorname{sgn}(\beta)$；$\beta = 0$ 时，$s \in [-1, +1]$。第 5 章会讨论凸优化的一般理论，即任何使次梯度方程（2.22）为零的解 $(\hat{\beta}, \hat{s})$ 都是原问题（2.9）的最优解，其中，$\hat{s} \in \operatorname{sgn}(\hat{\beta})$。

因此，求解式（2.22）可得到式（2.10）和式（2.11）的解。

习题 2.4　求证：问题式（2.5）的次梯度方程具有与式（2.6）相同的形式，并由此推导出坐标下降步骤中的式（2.14）和式（2.15）。

习题 2.5　**lasso 拟合值的唯一性**。对于 $\lambda \geqslant 0$，假设有两个 lasso 解 $\hat{\beta}$ 和 $\hat{\gamma}$，对应的最优解的值为 c^*。

(a) 求证：必须要 $\boldsymbol{X}\hat{\beta} = \boldsymbol{X}\hat{\gamma}$，即两个解产生相同的最优解值。（提示：如果这个等式不成立，则由函数 $f(\boldsymbol{u}) = \|\boldsymbol{y} - \boldsymbol{u}\|_2^2$ 的严格凸性和 ℓ_1 惩罚项的凸性可以推出矛盾。）

(b) 若 $\lambda > 0$，则有 $\|\hat{\beta}\|_1 = \|\hat{\gamma}\|_1$（Tibshirani 2013）。

习题 2.6　这里使用 bootstrap 作为 lasso 推理的基础。

(a) 对于犯罪数据，采用 bootstrap 估计 lasso 得到的系数的标准误差（如表 2-2 中间部分所示）。使用非参数 bootstrap 时，从训练数据中有放回抽取特征和输出值 (x_i, y_i)，让来自原 lasso 拟和所估计的 t 值不变。估计得到全为零的系数的概率。

(b) 重复（a），但对每个重复的 bootstrap，要重新估算 $\hat{\lambda}$。将结果与（a）的结果进行比较。

习题 2.7　用最小二乘来拟合有 k 特征的线性模型。求证：其自由度 [见式（2.17）] 等于 k。

习题 2.8　**在正交情形下 lasso 的自由度。** 假设 $y_i = \beta_0 + \sum_j x_{ij}\beta_j + \varepsilon_i$，其中 $\varepsilon_i \sim N(0,\sigma^2)$，$x_{ij}$ 是固定的（非随机）。假设对特征进行了中心化，并假定它们不相关，即对所有的 i 和 k，有 $\sum_i x_{ij}x_{ik} = 0$。Stein 引理（Stein 1981）指出：对于 $Y \sim N(\mu,\sigma^2)$ 和所有绝对连续函数 g（即 $\mathbb{E}|g'(Y)| < 0$）有

$$\mathbb{E}(g(Y)(Y-\mu)) = \sigma^2\mathbb{E}(g'(Y)) \tag{2.23}$$

使用式（2.32）证明，在正交情形下 lasso 的自由度式（2.17）等于 k，即为估计的非零系数个数。

习题 2.9　求证式（2.19）的解为式（2.20）。

习题 2.10　**从健壮回归的角度考虑 lasso。** 考虑标准线性回归问题的健壮性版本，其中保护特征免于扰动。为了完成这样的功能，需要最小–最大化下面的目标函数：

$$\underset{\beta}{\text{minimize}}\max_{\boldsymbol{\Delta}\in\mathcal{U}}\left\{\|\boldsymbol{y}-(\boldsymbol{X}+\boldsymbol{\Delta})\beta\|_2^2\right\} \tag{2.24}$$

其中，扰动 $\boldsymbol{\Delta} := (\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_p)$ 属于 $\mathbb{R}^{N\times p}$ 中的如下子集：

$$\mathcal{U} := \left\{(\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\ldots\boldsymbol{\delta}_p) \mid \|\boldsymbol{\delta}_j\|_2 \leqslant c_j \quad j=1,2,\ldots,p\right\} \tag{2.25}$$

因此，每个特征值 x_{ij} 被扰动的最大量为 c_j，即特征的扰动向量的 ℓ_2 范数要小于等于 c_j。用于不同特征的扰动也彼此独立。我们要在“最差”可允特征扰动下寻找最小误差平方和的系数。假设 $\boldsymbol{y}$ 和 $\boldsymbol{X}$ 的列已经归一化，而且没有截距。

求证这个问题的解等于

$$\min_{\beta\in\mathbb{R}^p}\left\{\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2 + \sum_{j=1}^{p} c_j|\beta_j|\right\} \tag{2.26}$$

若 $c_j = \lambda, j = 1,2,\ldots,p$，则可得到 lasso。因此式（2.26）可被看作一种用于防止测量特征值产生不确定性的方法，过多的不确定性会导致更大的收缩量，参见 Xu，Caramanis and Mannor（2010）。

习题 2.11　**健壮性回归与约束优化。** 本习题并不涉及 lasso 本身，而是关注回归中与 ℓ_1 范数相关的用途。设有如下模型：

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \gamma_i + \varepsilon_i$$

其中 $\varepsilon_i \sim N(0,\sigma^2)$ 和 $\gamma_i(i=1,2,\ldots,N)$ 都是未知常量。

令 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_N)$，则会有如下的优化问题：

$$\underset{\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^N}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \gamma_i)^2 + \lambda \sum_{1}^{N} |\gamma_i| \tag{2.27}$$

此处思路为：对于每个 i，γ_i 允许 y_i 为离群值，若 $\gamma_i = 0$ 表示观测值中没有噪声。惩罚项有效限制了噪声的数量。

(a) 求证：这个问题同时关于 β 和 $\boldsymbol{\gamma}$ 的凸函数。

(b) 思考如下 Huber 损失函数。

$$\rho(t; \lambda) = \begin{cases} \lambda |t| - \lambda^2/2, & |t| > \lambda \\ t^2/2, & |t| \leqslant \lambda \end{cases} \tag{2.28}$$

这是一个锥形平方误差损失，对 $|t| \leqslant \lambda$ 部分是二次函数，但超出该范围为线性，这样可以减少估计 β 时噪声的影响。若尺度参数 σ 设为 1，可求解下面的目标函数来得到 Huber 健壮回归：

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \sum_{i=1}^{N} \rho(y_i - \sum_{j=1}^{p} x_{ij}\beta_j; \lambda) \tag{2.29}$$

求证：式（2.27）和式（2.29）有相同的解 $\hat{\beta}$（Antoniadis 2007, Gannaz 2007, She and Owen 2011）。

第 3 章　广义线性模型

第 2 章只介绍了如何用最小二乘来拟合线性回归模型，这种线性模型适用于输出变量为连续值、误差为高斯分布的情形。但是，在实际应用中还有其他类型的输出变量。例如，二值变量就常用于表示某些特征为“有”或“无”，如生物学中鉴定是“癌”细胞还是“正常”细胞，网页浏览分析中的“点击”和“没点击”；这种情况采用二项式分布更加合适。有时候，输出变量为计数方式，例如排队的顾客数量、检测到的光子数量，这可能服从泊松分布。本章将讨论简单线性模型的推广形式，以及适用于这些模型的 lasso 方法。

3.1　引　　言

线性逻辑斯蒂模型（linear logistic model）在实际应用中经常用到，其输出变量的形式为二值编码 $Y \in \{0,1\}$，其建模思路为：将对数似然比率作为线性组合

$$\log \frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)} = \beta_0 + \beta^{\mathrm{T}} x \tag{3.1}$$

其中 $X = (X_1, X_2, X_p)$ 是输入向量，$\beta_0 \in \mathbb{R}$ 是截距，$\beta \in \mathbb{R}^p$ 是线性回归系数的向量。将上式变型，可得到一个条件概率的表达式

$$\Pr(Y=1|X=x) = \frac{\mathrm{e}^{\beta_0+\beta^{\mathrm{T}}x}}{1+\mathrm{e}^{\beta_0+\beta^{\mathrm{T}}x}} \tag{3.2}$$

通过观察可知：不管参数 (β_0, β) 取什么值，该模型的取值范围都是 (0,1)，这正好是概率的取值范围。逻辑斯蒂模型通常使用极大似然估计来拟合。

式（3.1）称为条件概率的 logit 变换函数，它是一个联接函数（link function）。通常，联接函数从条件期望 $\mathbb{E}(Y|X=x)$ 变换而来（这里是 $Y=1$ 的条件概率），这种变换能够使期望值更自然地缩放，从而不需要加约束就能拟合参数。另一个例子是：如果输出变量 Y 为计数值，其取值范围是$\{0, 1, 2, \dots\}$，则需要确保条件概率为正值。这里自然会采用对数线性模型

$$\log \mathbb{E}[Y|X=x] = \beta_0 + \beta^{\mathrm{T}} x, \tag{3.3}$$

该模型采用了对数联接函数，可以通过最大化数据的泊松对数似然函数来拟合。

模型（3.1）和（3.3）都属于**广义线性模型**（McCullagh and Nelder 1989）的特例。这些模型用指数族分布中的函数来刻画输出变量，这些指数族分布包括二项式

分布、泊松分布、高斯分布等。经转换后，输出变量的均值 $\mathbb{E}[Y|X=x]$ 就可以用线性模型来近似。更具体地讲，若用 $\mu(x)=\mathbb{E}[Y|X=x]$ 表示在给定 $X=x$ 下 Y 的条件期望，那么广义线性模型的形式为

$$g[\mu(x)]=\underbrace{\beta_0+\beta^{\mathrm{T}}x}_{\eta(x)} \tag{3.4}$$

其中 $g:\mathbb{R}\to\mathbb{R}$ 是一个严格单调的联接函数。例如，若输出变量 $Y\in\{0,1\}$，对于逻辑斯蒂回归模型，其 $\mu(x)=\Pr[Y=1|X=x]$，$g(\mu)=\mathrm{logit}(\mu)=\log(\mu/(1-\mu))$。而当输出变量属于高斯分布时，$\mu(x)=\beta_0+\beta^{\mathrm{T}}x$，$g(\mu)=\mu$ 即是一个标准的线性模型，这在第 2 章已经讨论过。

广义线性模型也可以用于实践中常常碰到的多分类问题，比如手写数字分类、语音识别、文本识别和癌症分类。多项式分布取代了二项式分布，就得到多分类逻辑斯蒂回归

$$\Pr[Y=k|X=x]=\frac{\mathrm{e}^{\beta_{0k}+\beta_k^{\mathrm{T}}x}}{\sum_{\ell=1}^{K}\mathrm{e}^{\beta_{0\ell}+\beta_\ell^{\mathrm{T}}x}} \tag{3.5}$$

每个变量共有 K 个系数（每个类一个）。

本章讨论拟合广义线性模型的方法，这些模型会最小化带有 ℓ_1 惩罚项的极大似然估计或与之等价的极小负对数似然估计

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{-\frac{1}{N}\mathcal{L}(\beta_0,\beta;\boldsymbol{y},\boldsymbol{X})+\lambda\|\beta\|_1\right\} \tag{3.6}$$

在这里，$\boldsymbol{y}$ 是 N 维输出向量，$\boldsymbol{X}$ 是 $N\times p$ 维的输入矩阵，而对数似然函数 $\mathcal{L}$ 的具体形式会因广义线性模型的不同而不同。在标准线性模型中，输出变量为高斯分布，则 $\mathcal{L}(\beta_0,\beta;\boldsymbol{y},\boldsymbol{X})=\frac{1}{2\sigma^2}\|\boldsymbol{y}-\beta_0\mathbf{1}-\boldsymbol{X}\beta\|_2^2+c$，其中 c 是一个独立于 (β_0,β) 的常数，这时式（3.6）中的优化问题就是一个普通的线性最小二乘 lasso。

对于其他的相关模型，也可采用类似的 ℓ_1 正则化形式。比如生存模型（survival model），它的输出为死亡的时间，对象消失后可能要删除。在这种情况下，通常会用 Cox 比例风险（hazard）模型，其形式为

$$h(t|x)=h_0(t)\mathrm{e}^{\beta^{\mathrm{T}}x} \tag{3.7}$$

对每个变量 x 而言，$t\to h(t|x)$ 为一个**风险函数**：$h(t|x)$ 表示在 $Y=t$（t 为给定的生存时间）时的失败概率。函数 h_0 为基准风险函数，对应 $x=0$。

另一个例子是机器学习中的支持向量机（Support-Vector Machine, SVM），这是一个十分常见的分类器，其输出值为 $y\in\{+1,-1\}$，[①]其最简单形式为 $f(x)=$

① 对于 SVM，符号函数可以很方便地对二值输出变量进行编码。

$\beta_0 + \beta^{\mathrm{T}}x$，这是一个线性分类器，其中预测类由 $\mathrm{sgn}(f(x))$ 给出。因此，通过判断间隔 $yf(x)$ 是否为正，就可知道分类是否正确。传统的软间隔线性 SVM 通过求解以下优化问题 [①]得到

$$\underset{\beta_0,\beta}{\mathrm{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \underbrace{[1 - y_i f(x_i)]_+}_{\phi(y_i f(x_i))} + \lambda ||\beta||_2^2 \right\} \tag{3.8}$$

第一项即铰合损失函数，用于惩罚负间隔，即惩罚不正确的分类。通常，对于标准线性 SVM（3.8）而言，最优解向量 $\beta \in \mathbb{R}^p$ 并不稀疏，因为二范数并没有强制稀疏（sparsity-enforcing）的性质。但若用 ℓ_1 范数 $||\beta||_1$ 替代二范数（即 ℓ_1 线性 SVM），就会得到稀疏解。

接下来的内容会详细讨论这几种方法。每种方法都将提供具体的应用例子，并讨论出现的问题和拟合模型的计算方法。

3.2　逻辑斯蒂回归模型

在半个世纪以前，逻辑斯蒂回归模型便常常应用于生物医学研究，目前它在数据建模中的应用更为广泛。在高维环境，即特征数 p 远大于样本数 N 的情况下，逻辑斯蒂回归模型不能直接使用。当 $p > N$ 时，任何线性模型都会出现参数过多的情况。此时为了拟合得到一个更稳定的模型，必要进行正则化。这种高维模型有着广泛的应用。比如，在文本分类中，其特征通常取两个值（“出现”或“没出现”），对于一个预先定义好的字典，相应的 p=20 000，有时甚至更多。另一个例子是全基因组关联研究（Genome-Wide Association Study, GWAS），所涉及基因类型至少会有 $p = 500\,000$ 个 SNP，而输出变量为是否有某种疾病。SNP（读作 “snip”）是指单核苷酸多态性，通常用三个值{AA, Aa, aa}表示三种可能出现的状态，其中 A 表示野生型（wild-type），而 a 表示变异型（mutation）。

当输出变量为二值类型时，通常会采用 0/1 编码方式。接下来要估计条件概率 $\Pr(Y = 1|X = x) = \mathbb{E}[Y|X = x]$。式（3.1）的逻辑斯蒂模型是带有 ℓ_1 正则化的负对数似然函数，即

$$\begin{aligned} & -\frac{1}{N} \sum_{i=1}^{N} \{ y_i \log \Pr(Y = 1|x_i) + (1 - y_i) \log \Pr(Y = 0|x_i) \} + \lambda ||\beta||_1 \\ = & -\frac{1}{N} \sum_{i=1}^{N} \{ y_i(\beta_0 + \beta^{\mathrm{T}} x_i) - \log(1 + \mathrm{e}^{\beta_0 + \beta^{\mathrm{T}} x_i}) \} + \lambda ||\beta||_1 \end{aligned} \tag{3.9}$$

① 这不是引出支持向量机的标准方式，3.6 节会更详细地讨论这个话题。

在机器学习领域，输出变量 Y 的值通常带有正负号（如{−1，+1}），而不是{0,1}；当输出值是这种形式时，则带惩罚项的（负）对数似然形式为

$$\frac{1}{N}\sum_{i=1}^{N}\log(1+\mathrm{e}^{-y_i f(x_i;\beta_0,\beta)})+\lambda||\beta||_1 \tag{3.10}$$

其中 $f(x_i;\beta_0,\beta):=\beta_0+\beta^{\mathrm{T}}x_i$。对给定的一对输入和输出变量 (x,y)，$yf(x)$ 表示**间隔**（margin）：间隔值为正数表示函数对样本进行了正确的分类，间隔值为负数表示函数对样本进行了错误的分类。由式（3.10）可知，最大化似然函数其实是最小化一个损失函数，该损失函数会随间隔增加而单调递减。3.6.1 节会讨论间隔与惩罚项之间的影响。

3.2.1 示例：文本分类

ℓ_1 正则化逻辑斯蒂回归已然广泛应用于文本分类，这里我们用 20 个新闻组数据集（Lang 1995）来介绍基于 ℓ_1 正则化的逻辑斯蒂回归。特征集和分类情况可参考 Koh、Kim and Boyd（2007）[①]给出的定义。数据集中有 N=11 314 个文档，p=777 811 个特征，其中 52%为正样本。所有文本只有 0.05%的非零特征。

图 3-1 用 R 的 glmnet 包绘制而成。图中的曲线由 100 组 λ 值计算得到，这些 λ 值之间的间隔在 log 尺度下一样。可以按**可解释偏差比**（fraction deviance explained）[②③]

$$\mathrm{D}_\lambda^2=\frac{\mathrm{Dev}_{\mathrm{null}}-\mathrm{Dev}_\lambda}{\mathrm{Dev}_{\mathrm{null}}} \tag{3.11}$$

对训练集上的解进行排序（index）。其中，偏差 Dev_λ 定义为，带有正则参数 λ 的对数似然模型减去“饱和”模型的差值的平方。$\mathrm{Dev}_{\mathrm{null}}$ 是指空模型的偏差[④]，空模型指常数（均值）模型。这些数据为可分数据，因此需要考虑 λ 的选择范围，不要使所拟合的模型太过靠近饱和模型（饱和模型没有系数，详见 3.2.2 节）。

模型中非零系数的最大个数可以通过 min（N,p）得到，本例的非零系数个数为 11 314。在图 3-1 中，glmnet 没有得到最后的解，所以非零系数最多的模型也只有 5 277 个非零系数。虽然将图按 $\log(\lambda)$ 或者 $||\widehat{\beta}(\lambda)||_1$ 来画会更加自然些，但这两种方式都将面对 $p\gg N$ 的问题。前者依赖于数据和具体问题，而且不能表示出过拟合的数量；而对于后面的度量，图的右半部分被无关紧要的值占据，此处的系数及相应范数会激增。

---

① 正样本包括 10 组数据，命名形式为 sci.*、comp.* 和 misc.forsale，其余的为负样本。特征集为三元模型，忽略了文本标题、停用词表和只在单个文本中出现的特征。

② D^2 和 R^2 概念相近，R^2 是指回归中可解释方差比。

③ 可解释方差比也可称为可决系数。——译者注

④ 偏差也可称为残差。——译者注

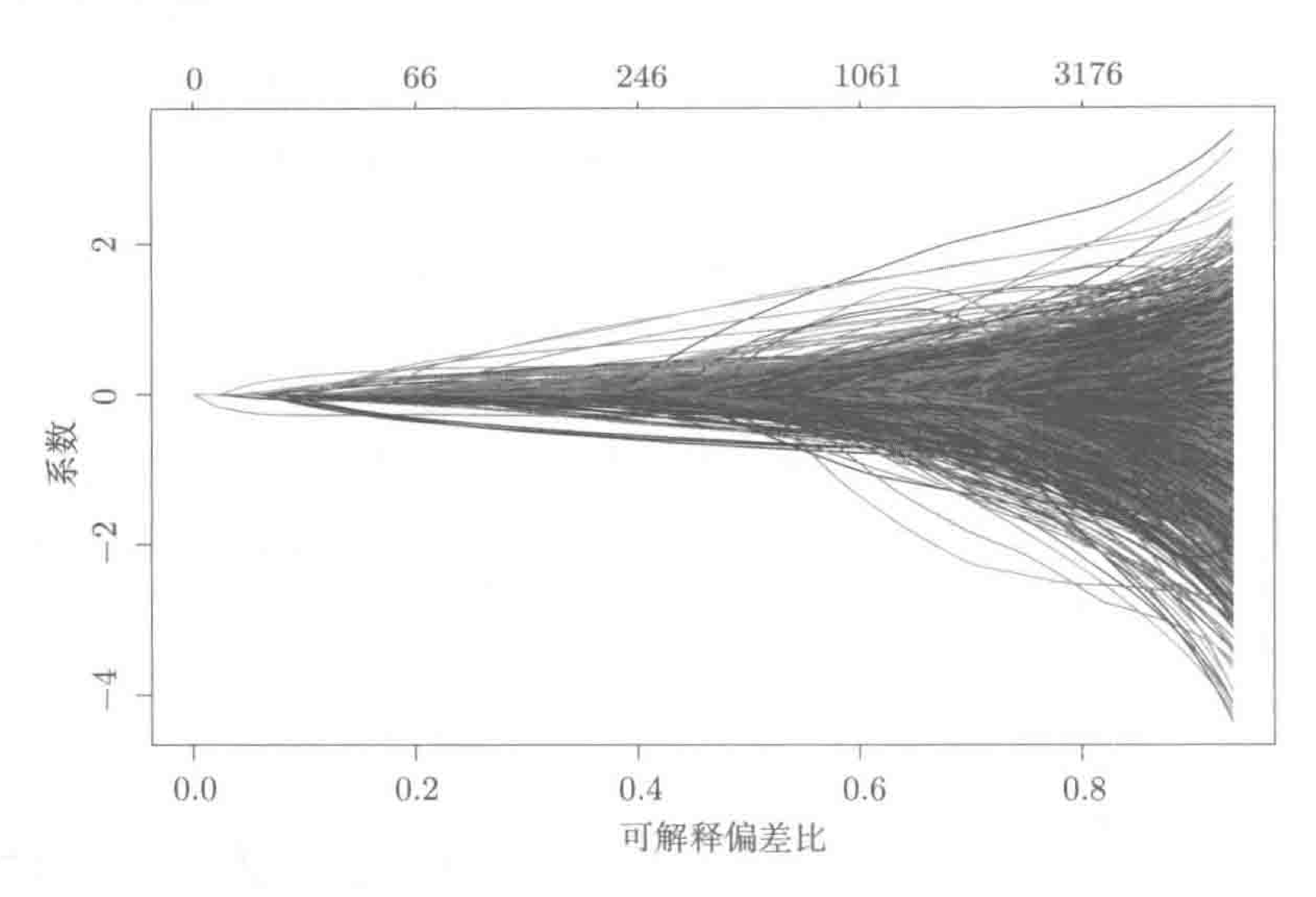


图 3-1　ℓ_1 正则化逻辑斯蒂回归应用于“新闻组”数据，进行文本分类的系数图。这里将所有的文本大致分为两类，共 0.78M 特征，其中只有 0.05%的特征是非零的。图中系数表示为可解释的零偏差分数的函数

图 3-2 给出了数据的十重交叉验证结果及训练误差率。图中数据依然按照训练数据上的可解释偏差比排序。图 3-3 为岭回归结果，与图 3-2 类似。图 3-3 中交叉验证的错误率也与 lasso 回归相似。两个模型的非零系数个数为 p=777 811，而图 3-2 中的非零系数最多为 5277。但是，在本例中，岭回归的非零系数为 11 314，这个值等于 $\min(N,p)$，同 lasso 差别不大。从计算角度来说，岭回归更加耗时。图 3-3 用 glmnet 包计算岭回归，十重交叉验证耗时 8.3 分钟，而 lasso 耗时 1 分钟。另外，也可以用核技巧计算岭回归（Hastie and Tibshirani 2004），但这需要对 11 314×11 314 矩阵进行奇异值分解或其他类似的分解。

本例在 2.6 GHz Macbook Pro 上通过 glmnet 求解图 3-1 中 100 个不同 λ 值所对应的模型，总共耗时 5 秒。这个例子有如此多的特征，筛选特征就可以显著提升计算速度。例如，正则化拟合过程中选出的第一个特征使得 $\lambda_{\max} = \max|\langle x_j, \boldsymbol{y}-\overline{\boldsymbol{p}}\rangle|$，式中 $\boldsymbol{y}$ 是二值输出向量，$\overline{\boldsymbol{p}} = \mathbf{0.521}$ 是均值向量。这是 λ 的初始值，这个值是使所有系数值都为零的最小值。在计算正则化模型的过程中，当 λ 值从 $\lambda_{\max}$ 下降到稍小的值 λ_1 时，就可以筛除出大部分变量，这些变量的内积远远小于 λ_1。用较小的子集计算模型的解，就能够检验出删除的特征变量是否在错误率中被忽略。求解的过程可以重复运用特征变量和当前残差的内积。在计算上面这个例子时，glmnet 包运用了这种“强规则”（strong rule）筛选法。5.10 节将详细介绍强规则和其他加速计算方法。

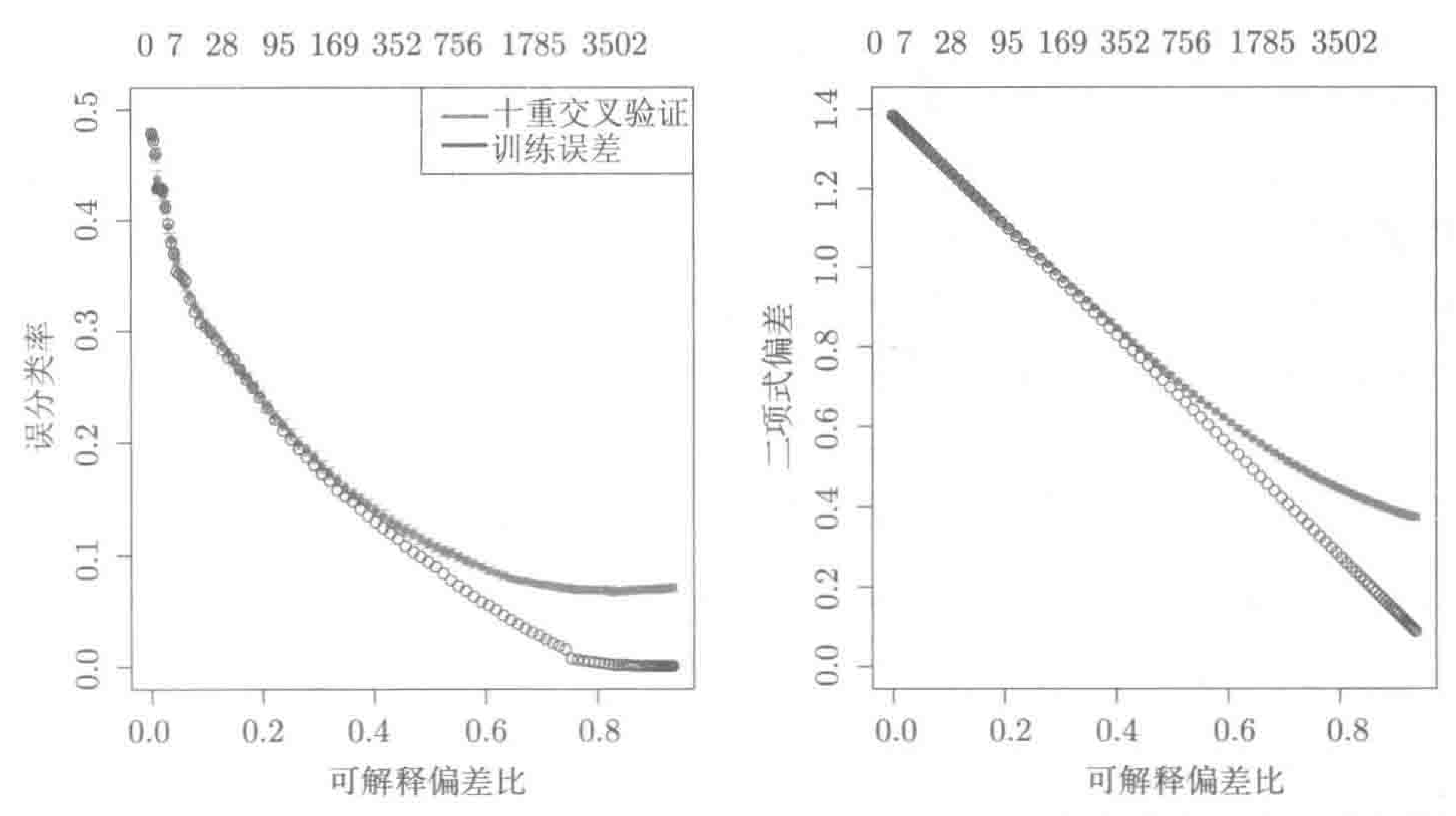


图 3-2 基于 lasso（ℓ_1）惩罚的逻辑斯蒂回归。红线表示新闻组数据的十重交叉验证结果，以及每一点的标准差（在图中并不明显）。左图为误分类率，右图为偏差。图中蓝线为相应的训练误差。图中上侧数字为每个模型的非零参数个数（见彩插）[①]

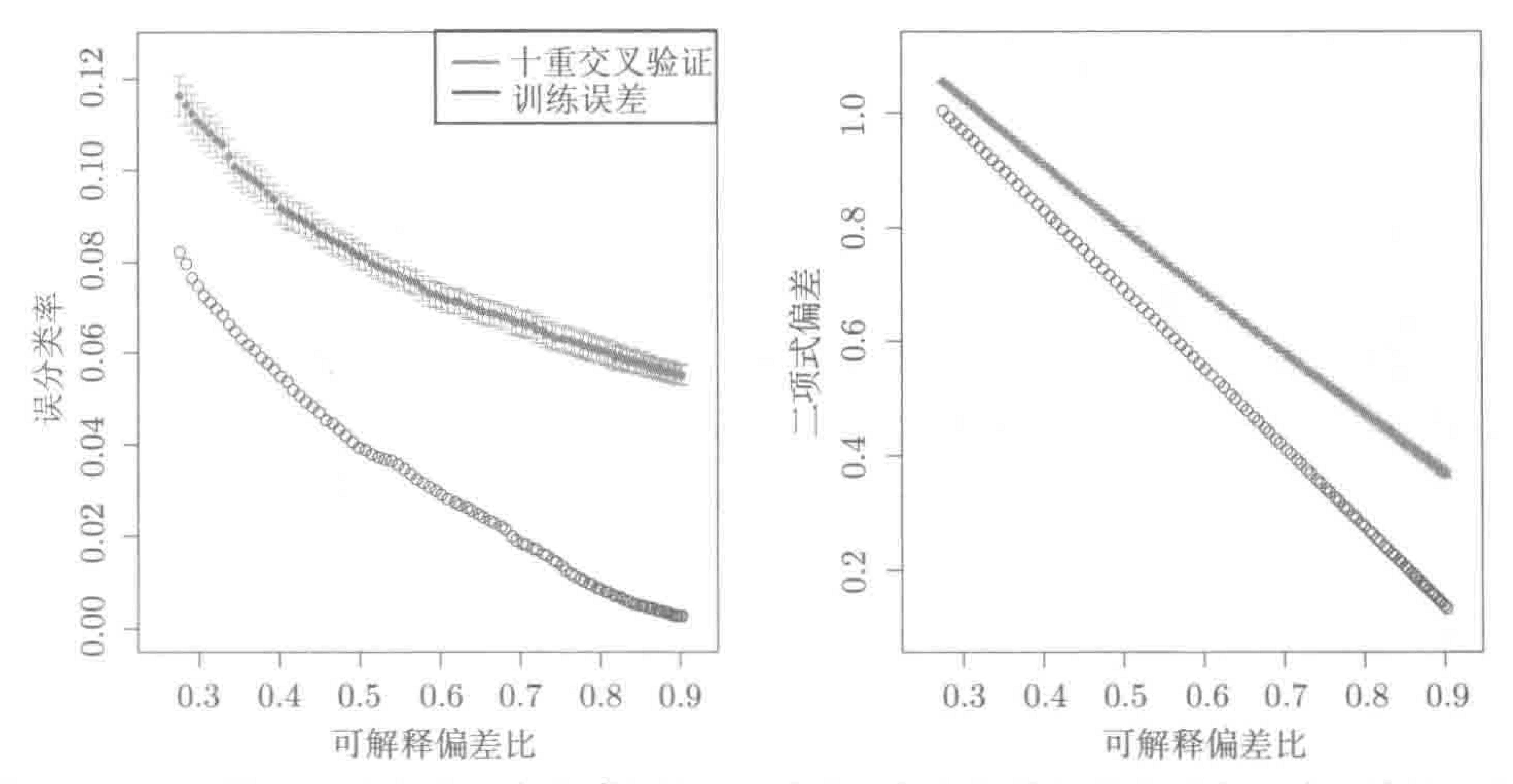


图 3-3 基于岭（ℓ_2）惩罚的逻辑斯蒂回归：红线表示新闻组数据的十重交叉验证结果，以及每一点的标准差范围。左图为误分类率，右图为偏差。图中蓝线为相应的训练误差（见彩插）

3.2.2 算法

二分类逻辑斯蒂回归是线性回归的常见推广形式，因此下面会主要介绍基于

① 本书所有图的彩色版请到图灵社区本书主页下载，www.ituring.com.cn/book/1723。—— 编者注

lasso 惩罚的逻辑斯蒂回归模型。式 (3.9) 是凸函数，似然部分可微，因此理论上可将此问题看成一个标准的凸优化问题 (Koh et al. 2007) 来求解。

坐标下降法对于此类问题十分有效，在参考文献中我们列出了与该方法相关的大量研究文献，也可参见 2.4.2 节和 5.4 节。glmnet 包用了一种近点牛顿 (proximal-Newton) 迭代方法，通过一个二次函数重复迭代逼近负对数似然函数 (Lee, Sun and Saunders 2014)。具体来说，由当前的 $(\widetilde{\beta}_0, \widetilde{\beta})$ 估计值可得到二次方程

$$Q(\beta_0, \beta) = \frac{1}{2N} \sum_{i=1}^{N} w_i (z_i - \beta_0 - \beta^{\mathrm{T}} x_i)^2 + C(\widetilde{\beta}_0, \widetilde{\beta}) \tag{3.12}$$

其中 C 是一个与 $(\widetilde{\beta}_0, \widetilde{\beta})$ 无关的常量，并且

$$z_i = \widetilde{\beta}_0 + \widetilde{\beta}^{\mathrm{T}} x_i + \frac{y_i - \widetilde{p}(x_i)}{\widetilde{p}(x_i)(1 - \widetilde{p}(x_i))}, \quad w_i = \widetilde{p}(x_i)(1 - \widetilde{p}(x_i)) \tag{3.13}$$

其中，$\widetilde{p}(x_i)$ 是 $\Pr(Y = 1|X = x_i)$ 的当前估计值。每一次外部循环都意味着求解一个加权的 lasso 回归。可以在一个 λ 值的细网格上执行热启动，只需要几次外部循环迭代即可达到最优，因为此时局部二阶近似已经很好了。3.7 节和 5.4.2 节将讨论 glmnet 包的特性。

3.3 多分类逻辑斯蒂回归

有些分类问题的类别数 K 大于 2。在机器学习领域，常用方法就是建立全部 $\binom{K}{2}$ 个分类器 (一对一，或 OvO)，然后将样本归类到可能性最大的类别中。另一种方法是建立一对多 (OvA) 分类器，将某一类视为正样本，其余的全当成负样本。这两种方法都有坚实的理论基础，也都有局限性。OvO 方法会浪费计算资源，而 OvA 方法会带入屏蔽作用 (Hastie et al. 2009, Chapter 4)。多分类逻辑斯蒂回归能够引入一种更加自然的方法。这里可采用多项式似然函数，用指数线性表达式来表示概率

$$\Pr(Y = k|X = x) = \frac{\mathrm{e}^{\beta_{0k} + \beta_k^{\mathrm{T}} x}}{\sum_{\ell=1}^{K} \mathrm{e}^{\beta_{0\ell} + \beta_\ell^{\mathrm{T}} x}}. \tag{3.14}$$

这是一个冗余模型，因为对每一类中的线性模型加上一个线性项 $\gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}} x$，概率并不会变化。因此，习惯做法是将其中一类模型 (一般是最后一类) 参数设为零，这样需估计的模型就只有 $K - 1$ 个线性函数了。选择不同的基类 (base class)，用最大化似然函数拟合得到的模型是一样的，因为不同模型的参数之间具有等变性 (equivariant)。(一个基类的参数的解也可从另一种基类的参数解中推导出来。)

下面采用冗余但对称的方法来获得式 (3.14) 的参数，原因有两条：

- 需要对参数进行正则化，而在不同基类下，对参数正则化后的解不再具有等变性；
- 正则化方法会自动消除冗余（详解如下）。

对于样本 $\{(x_i, y_i)\}_{i=1}^N$，采用负对数似然函数的正则化形式为

$$-\frac{1}{N}\sum_{i=1}^{N}\log\Pr(Y=y_i|x_i;\{\beta_{0k},\beta_k\}_{k=1}^K)+\lambda\sum_{k=1}^{K}||\beta_k||_1 \tag{3.15}$$

指示响应变量用矩阵 $\boldsymbol{R}$ 表示，该矩阵大小为 $N\times K$，其元素 $r_{ik}=\mathbb{I}(y_i=k)$。可以将式（3.15）中似然函数部分重写为

$$\frac{1}{N}\sum_{i=1}^{N}w_i\left[\sum_{k=1}^{K}r_{ik}(\beta_{0k}+\beta_k^{\mathrm{T}}x_i)-\log\left\{\sum_{k=1}^{K}\mathrm{e}^{\beta_{0k}+\beta_k^{\mathrm{T}}x_i}\right\}\right] \tag{3.16}$$

每一个样本对应一个权重值 w_i，默认值为 $w_i=1/N$。这种形式便于对输出变量进行分组：每个样本 x_i 都对应一个集合，该集合含有 n_i 个多分类响应变量，其中 r_{ik} 表明 x_i 是否属于第 k 类。也就是说，矩阵 $\boldsymbol{R}$ 的每行为各类别比，这样 $w_i=n_i$ 为样本权重。

如前所述，对于每个特征 x_j，在 K 个参数偏移一个常数情况下，模型的概率及其对数似然不会变，即 $\{\beta_{kj}+c_j\}_{k=1}^K$ 和 $\{\beta_{kj}\}_{k=1}^K$ 会得出同样的概率。这样，c_j 的选择就取决于式（3.15）中的惩罚部分。显然，对于任意的候选解 $\{\widetilde{\beta}_{kj}\}_{k=1}^K$，最优的 c_j 需满足

$$c_j=\underset{c\in\mathbb{R}}{\arg\min}\left\{\sum_{k=1}^{K}\left|\widetilde{\beta}_{kj}-c\right|\right\} \tag{3.17}$$

因此，对于 j=1, …, p，式（3.17）的最大值为 $\{\widetilde{\beta}_{1j},\ldots,\widetilde{\beta}_{Kj}\}$ 的中值，见习题 3.3。因为截距 $\{\beta_{0k}\}_{k=1}^K$ 没有惩罚，所以需要处理其不确定性。在 glmnet 包中，要将它们的和约束为零。

3.3.1 示例：手写数字

这里以美国邮政局的手写数字数据（Le Cun, Boser, Denker, Henderson, Howard, Hubbard and Jackel 1990）为例进行讲解。数据集中有 N=7291 张数字{0,1,. . . ,9}的训练图像，所有图像都被数字化为 16×16 的灰度图。用 p=256 个像素作为特征，在此拟合一个 10 个类的 lasso 多分类模型。图 3-4 可以看作一个函数图像，即将一系列 λ 值作为变量，训练和测试误分类误差作为函数的输出。图 3-5 将系数展示在图上（平均约有 25%为 0）。高亮后，可以从图上看出对应的数字。

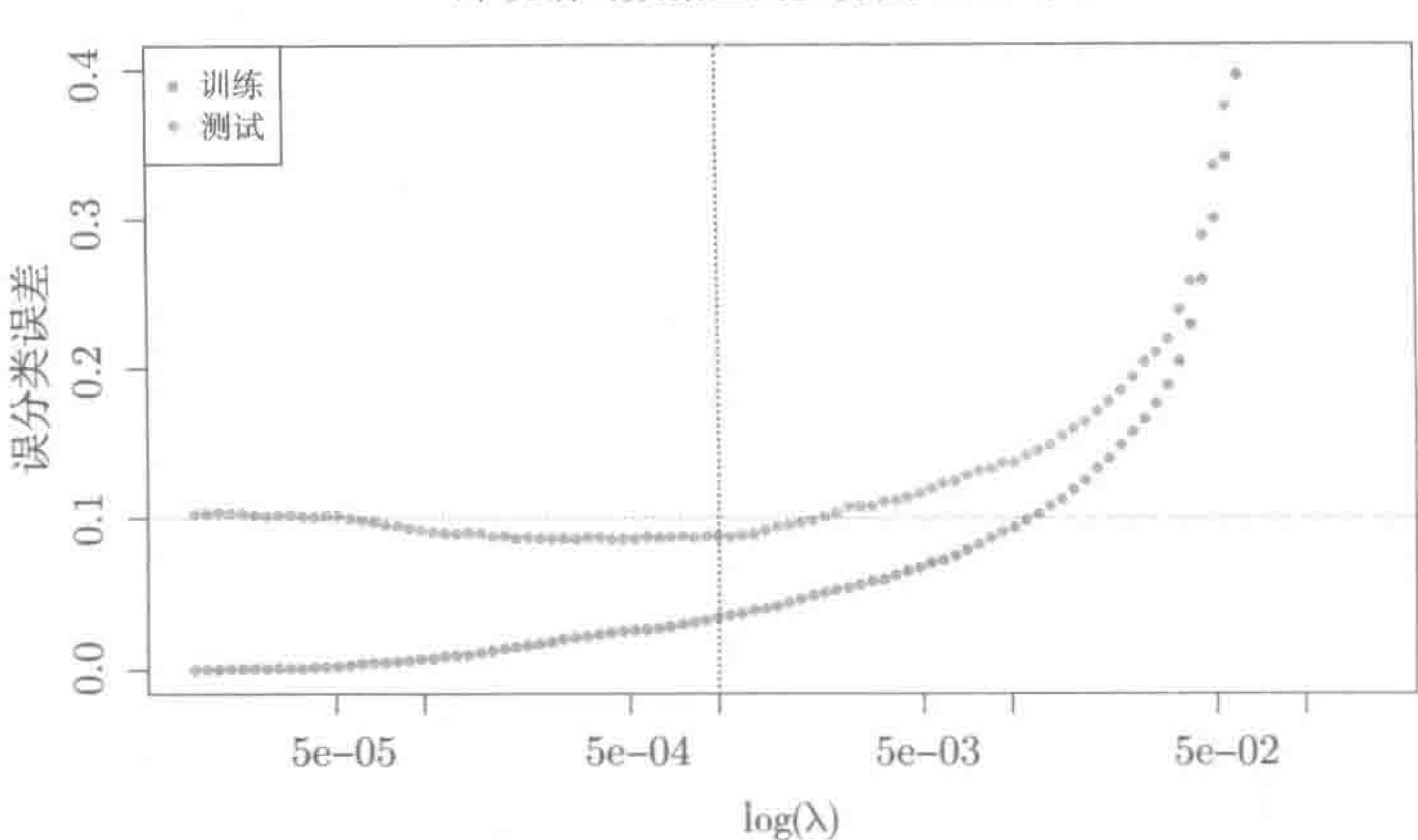


图 3-4　在邮政编码数据上，采用 lasso 模型所得到的训练误分类误差和测试分类误差，这是一个将 log(λ) 作为变量的函数图。图中最小的测试误差在 0.086 左右，最小的训练误差为 0。图中标出了 λ=0.001 时的值，此时每一类的系数如图 3-5 所示

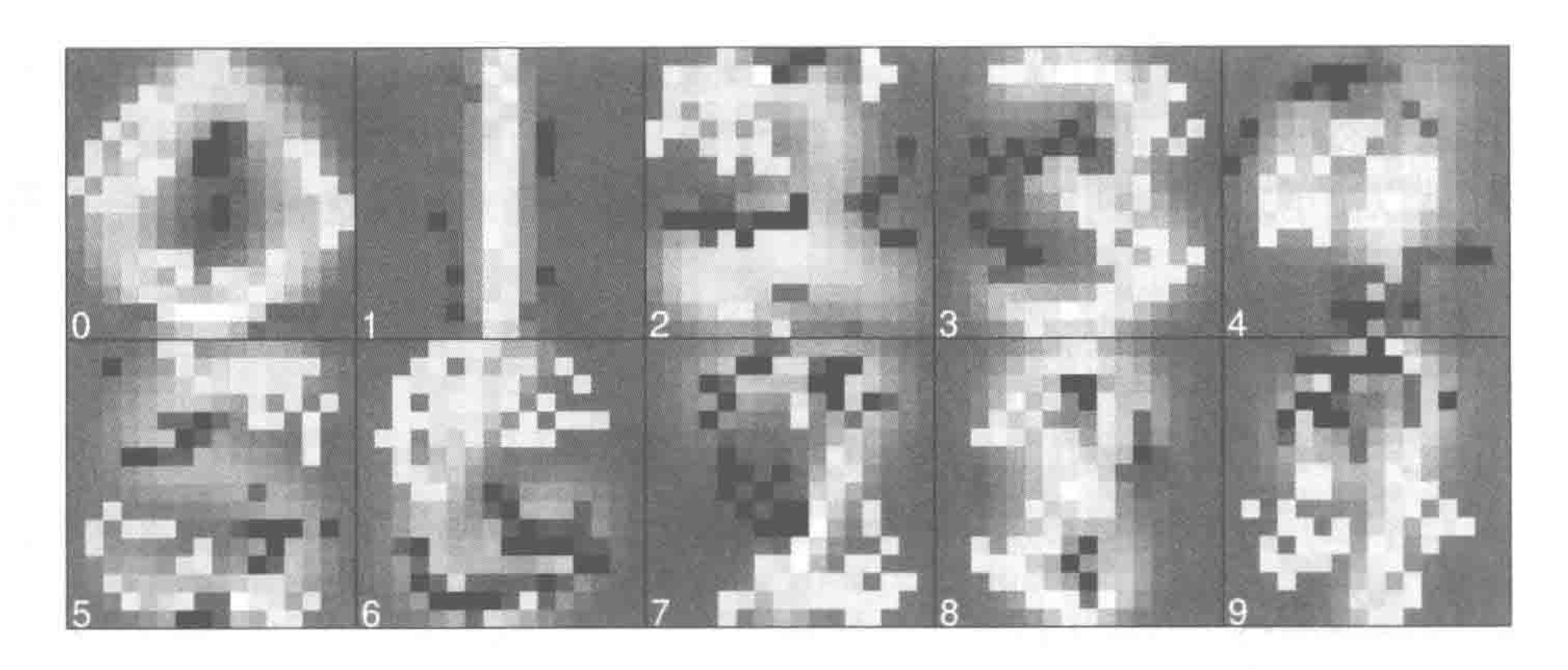


图 3-5　多类 lasso 回归中每一类数字的系数。灰色背景为每一类的平均训练样本。叠加在上的颜色（黄色为正系数，蓝色为负系数）表示为每一类的非零系数。注意，这些值为非零的地方有所不同，由此便产生了每一类的判别评价（score）。并非所有这些值都可解释（见彩插）

3.3.2　算法

尽管这个问题可以用标准凸优化软件求解，但是有研究发现坐标下降法更加有效（Friedman, Hastie, Simon and Tibshirani 2015）。在二分类问题中，外层循环是牛顿算法，内层循环是加权最小二乘。外部循环可以看作以当前估计值 $\{\widetilde{\beta}_{0k}, \widetilde{\beta}_k\}_{k=1}^K$ 为中心，对对数似然二次逼近。这里会采用同样的方法，只是在逼近时，仅改变其

中一类的参数，其余类的参数保持不变。具体而言，更新参数 $(\beta_{0\ell},\beta_\ell)$，会得到二次函数

$$Q_\ell(\beta_{0\ell},\beta_\ell) = -\frac{1}{2N}\sum_{i=1}^{N} w_{i\ell}(z_{i\ell}-\beta_{0\ell}-\beta_\ell^{\mathrm{T}}x_i)^2 + C(\{\widetilde{\beta}_{0k},\widetilde{\beta}_k\}_{k=1}^{K}) \tag{3.18}$$

C 表示独立于$(\beta_{0\ell},\widetilde{\beta}_\ell)$ 的常量，而

$$\begin{aligned} z_{i\ell} &= \widetilde{\beta}_{0\ell} + \widetilde{\beta}_\ell^{\mathrm{T}}x_i + \frac{y_{i\ell}-\widetilde{p}_\ell(x_i)}{\widetilde{p}_\ell(x_i)(1-\widetilde{p}_\ell(x_i))}, \\ w_{i\ell} &= \widetilde{p}_\ell(x_i)(1-\widetilde{p}_\ell(x_i)) \end{aligned}$$

其中 $\widetilde{p}_l(x_i)$ 是条件概率 $\Pr(Y=\ell|x_i)$ 的当前估计。这里采用的方法与二分类的情况类似，只是外部循环必须要以类别为基础。对于每个 λ 值，需要通过 $\ell \in \{1,\ldots,K\}$ 来创建外部循环，基于当前参数 $(\widetilde{\beta}_0,\widetilde{\beta})$ 计算部分二次逼近值 Q_ℓ。然后用坐标下降法求解加权 lasso 问题

$$\underset{(\beta_{0\ell},\beta_\ell)\in\mathbb{R}^{p+1}}{\text{minimize}} \{Q(\beta_{0\ell},\beta_\ell)+\lambda||\beta_\ell||_1\} \tag{3.19}$$

3.3.3 组 lasso 多分类

如图 3-5 所示，lasso 惩罚会在不同的类中选择不同的变量。也就是说，虽然每个类的系数向量都是稀疏的，但是整个模型的系数并不稀疏。在本例中，平均每个类有 25%的非零系数，而整个模型有 81%的变量系数为非零。

另一种方法是对一组系数 $\beta_j=(\beta_{1j},\beta_{2j},\ldots,\beta_{Kj})$ 采用组 lasso 惩罚（见 4.3 节），因此式（3.15）可以替换为

$$-\frac{1}{N}\sum_{i=1}^{N}\log\Pr(Y=y_i|X=x_i;\{\beta_j\}_{j=1}^{p}) + \lambda\sum_{j=1}^{p}||\beta_j||_2 \tag{3.20}$$

需要注意的是，约束条件中出现的是 ℓ_2 范数 $\|\cdot\|_2$ 的和，而非 ℓ_2 范数的平方和。这样就相当于对整个系数施加了一个 ℓ_1/ℓ_2 约束：对整个系数加上了 ℓ_2 范数的和。这种组惩罚的作用是让某一个变量的所有系数在或者不在模型中。在模型中时，这些系数全为非零（见习题 3.6），并且自动满足约束条件 $\sum_{k=1}^{K}\beta_{kj}=0$。式（3.20）是凸的，所以用标准的方法即可找到最优值。和之前一样，这里也可以采用坐标下降法，在每个系数向量 β_j 上采用块（block）坐标下降法，并保持其余的系数向量不变（见习题 3.7）。组 lasso 及其变种的相关知识可参见 4.3 节。

3.4 对数线性模型及泊松广义线性模型

当输出变量 Y 为非负且表示一个计数时，其均值为正，且用泊松似然进行推断。这种情形常用对数线性模型式（3.3）来保证其值为正。这里假设对每一个

$X=x$，输出变量 Y 服从泊松分布，其均值 μ 满足

$$\log\mu(x)=\beta_0+\beta^{\mathrm{T}}x \tag{3.21}$$

基于 ℓ_1 惩罚的负对数似然模型为

$$-\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0+\beta^{\mathrm{T}}x_i)-\mathrm{e}^{\beta_0+\beta^{\mathrm{T}}x_i}\}+\lambda||\beta||_1 \tag{3.22}$$

对于其他广义线性模型，这里可以使用迭代再加权最小二乘法拟合模型，这意味着要在每一次外部迭代中拟合一个加权 lasso 回归。同样，这里并不惩罚截距 β_0。容易看出，这里的约束是强制平均拟合值等于均值，即 $\frac{1}{N}\sum_{i=1}^{N}\widehat{\mu}_i=\bar{y}$，其中 $\widehat{\mu}_i:=\mathrm{e}^{\widehat{\eta}(x_i)}=\mathrm{e}^{\widehat{\beta}_0+\widehat{\beta}^{\mathrm{T}}x_i}$。

泊松模型通常用来建立比率（比如死亡率）模型。如果观察的时间段 T_i 各不相同，则计数均值为 $\mathbb{E}(y_i|X_i=x_i)=T_i\mu(x_i)$，其中 $\mu(x_i)$ 是单位时间间隔的比率。这个例子的模型为

$$\log(\mathbb{E}(Y|X=x,T))=\log(T)+\beta_0+\beta^{\mathrm{T}}x \tag{3.23}$$

观测的参数 $\log(T_i)$ 是偏移量，不需要拟合。后面的例中子也有偏移量。

分布平滑

泊松模型是估计分布的实用工具。下面的例子来自 Yoram Singer（Singer and Dubiner 2011）。假设由 N 个计数 $\{y_k\}_{k=1}^{N}$ 组成的样本服从 N 元多项式分布，令 $r_k=y_k/\sum_{l=1}^{N}y_\ell$ 为对应的比例向量。比如，在大型 Web 应用中，这些计数代表在给定的一周内美国每个州访问某一特定页面的人数。比例向量可能是稀疏的，这取决于具体的情况，因此可以朝一个更广、更稳定的分布 $\boldsymbol{u}=\{u_k\}_{k=1}^{N}$ 调整（例如，同样的人口，但时间跨度扩大到一年）。Singer and Dubiner（2011）提出了下面的问题。

$$\underset{q\in\mathbb{R}^N,\ q_k\geqslant 0}{\text{minimize}}\sum_{k=1}^{N}q_k\log\left(\frac{q_k}{u_k}\right),\quad ||\boldsymbol{q}-\boldsymbol{r}||_\infty\leqslant\delta,\quad \sum_{k=1}^{N}q_k=1 \tag{3.24}$$

总之，这里发现了一个分布，它与观察到的分布之差的 ℓ_∞ 范数会小于给定的 δ，用 Kullback–Leibler（KL）散度来衡量的话，与分布 $\boldsymbol{u}$ 很相近。式（3.24）优化问题的拉格朗日对偶形式为（见习题 3.4）

$$\underset{\beta_0,\boldsymbol{\alpha}}{\text{maximize}}\left\{\sum_{k=1}^{N}[r_k\log q_k(\beta_0,\alpha_k)-q_k(\beta_0,\alpha_k)]-\delta||\boldsymbol{\alpha}||_1\right\} \tag{3.25}$$

其中 $q_k(\beta_0,\alpha_k):=u_k\mathrm{e}^{\beta_0+\alpha_k}$。这就相当于拟合一个带偏移量 $\log(\mu_k)$ 的泊松广义线性模型。每个样本对应一个参数 α_k，以及非常稀疏的设计矩阵 $\boldsymbol{X}=\boldsymbol{I}_{N\times N}$。因

此，模型可以用稀疏矩阵方法有效拟合（见 3.7 节）。图 3-6 是一个模拟的例子，分布 u_k 是对连续分布（混合高斯分布）离散化后得到的。分布有 N=100 单元，总数为 $\sum_{k=1}^{N} y_k = 1000$ 的样本分布在这些单元上。如上所述，未惩罚的 β_0 确保了 $\sum_{k=1}^{N} \hat{q}_k = \sum_{k=1}^{N} \widehat{r}_k = 1$（见习题 3.5）。虽然图 3-6 只展示了一个解，但是解 $\widehat{q}_k(\delta)$ 在分布 u_k 和观测到的分布 r_k 之间的变化已经非常平滑。

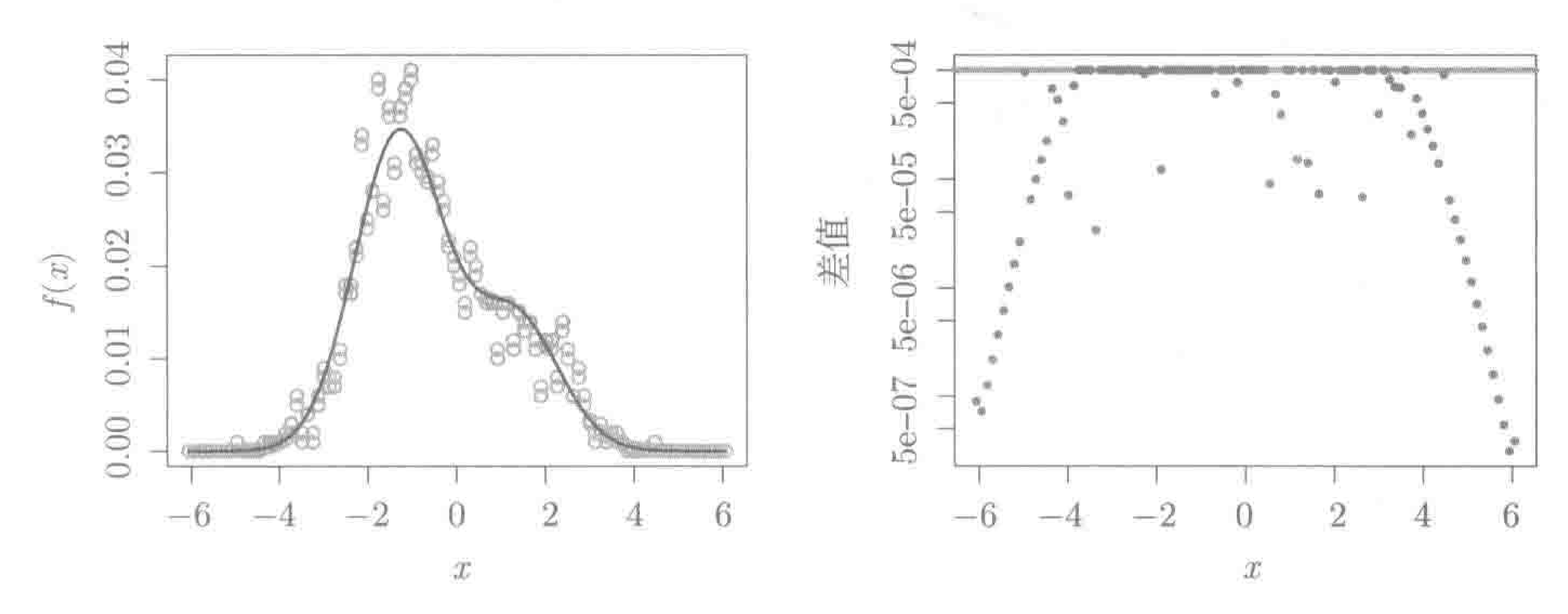


图 3-6　通过泊松模型估计分布。在左图中，黑实线是分布 $\boldsymbol{u}$，这里表示离散化一维分布 $f(x)$ 到 100 个单元。蓝点表示观测到的分布，黄点表示模型恢复后的分布。观测到的分布可能会有一些为零的计数，经过模型恢复后的分布与 $\boldsymbol{u}$ 支撑相同。右图为 N=100 时，$|\widehat{q}_k - r_k|$ 的差，其约束小于 $\delta = 0.001$，即水平黄线以下（见彩插）

3.5　Cox 比例风险模型

医学研究常常关注诊疗后的患者病故时间及康复时间。在治疗之后，患者会接受跟踪调查，有些患者在跟踪的过程中因搬迁而失去联系，有些则因其他原因而死亡。这些结果称为**右截尾**（right censored）数据。T 表示潜在的生存时间，对于每一个病人，观测得到 $Y = \min(T, C)$，其中 C 是**截尾时间**（censoring time）。重点通常在于存活函数 $S(t) := \Pr(T > t)$，即超过某一给定时间 t 的存活概率。

图 3-7 中的黑线表示对 N=240 例淋巴瘤患者（Alizadeh et al. 2000）的 $S(t)$ 估计。图中的尖刺代表截尾点，表示对某一病人不再跟踪调查的时刻。尽管所有的生存曲线都对这些数据进行了有效总结，但是加入协变量后能更好地对**风险函数**（即 S 函数的单调变换）建模。具体而言，t 时刻的风险为

$$h(t) = \lim_{\delta \to 0} \frac{\Pr(Y \in (t, t+\delta) | Y \geqslant t)}{\delta} = \frac{f(t)}{S(t)} \tag{3.26}$$

对应于 t 时刻的瞬时死亡率，即一直存活到 t 时刻。

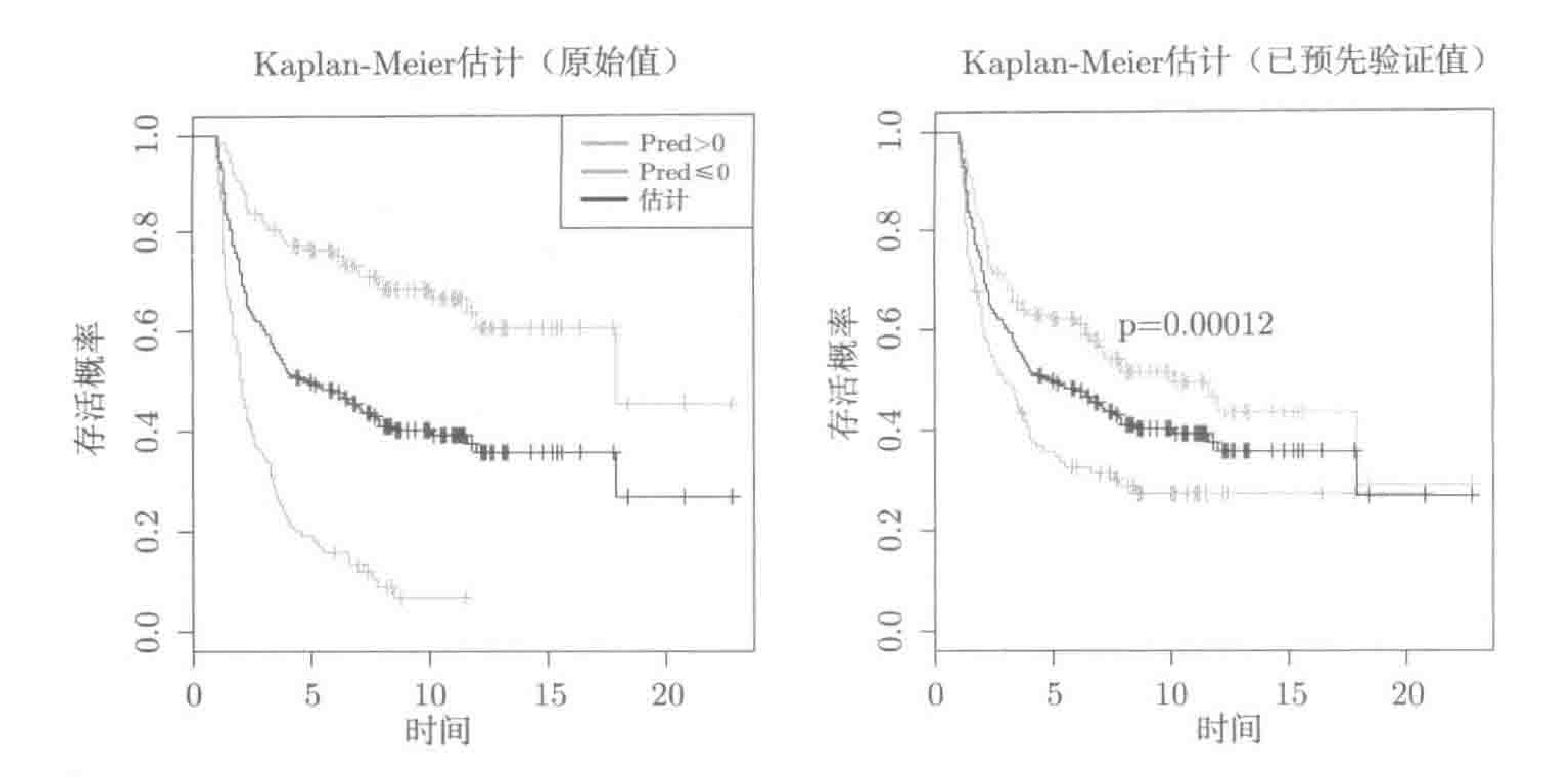


图 3-7　黑线为针对淋巴瘤数据的 $S(t)$ 函数的 Kaplan-Meier 估计。左图用 Cox 比例风险的 lasso 模型的预测值分割数据，并通过交叉验证进行选择。尽管模型参数经过了交叉验证选择，但预测值基于全部的训练数据已经极度优化了。右图在对整个数据集预测时先进行了预验证，去掉了训练集的偏差。尽管数据分割并不强，但依然十分明显。尖刺表示截尾时间。右图中的 p-value 由 log-rank 检验得出

现在来讨论产生图 3-7 中蓝色和黄色生存曲线的 Cox 比例风险模型。比例风险模型（CPH）基于风险函数

$$h(t,x) = h_0(t)\mathrm{e}^{\beta^{\mathrm{T}} x} \tag{3.27}$$

其中 $h_0(t)$ 是基准风险（各个样本在 x=0 时的风险）。

这里有形式为 (x_i, y_i, δ_i) 的数据，其中 δ_i 是一个二值指示变量，表示 y_i 是处于死亡时间还是截尾时间。淋巴瘤数据有 p=7399 个特征，均作为基因表达式的度量。在 N=240 个样本中，共有 102 个样本为右截尾。下面需要求解

$$\underset{\beta}{\text{minimize}} \left\{ -\sum_{\{i|\delta_i=1\}} \log\left[\frac{\mathrm{e}^{\beta^{\mathrm{T}} x_i}}{\sum\limits_{j\in R_i} \mathrm{e}^{\beta^{\mathrm{T}} x_j}}\right] + \lambda ||\beta||_1 \right\} \tag{3.28}$$

来拟合 ℓ_1 惩罚 CPH 模型，其中 R_i（$i = 1, \ldots, N$）为 y_i 时刻所研究的存活样本风险集。第一个参数是**偏似然**（partial likelihood）对数，对应在风险集中观测到死亡的条件概率。注意，基准风险没有起作用，这是该方法的一个特殊之处。这里假设样本之间没有关联，即存活时间是独立的。如果存在相互关联的情况，则偏似然需要调整。

图 3-8 是对淋巴瘤数据进行拟合后得到的模型系数。因为 $p \gg N$，所以当 $\lambda \downarrow 0$ 时，模型趋于“饱和”，即一些参数将会趋于 $\pm\infty$，对数偏似然函数会趋于 0。当 λ 变小时，这种不好的行为就会显现出来。

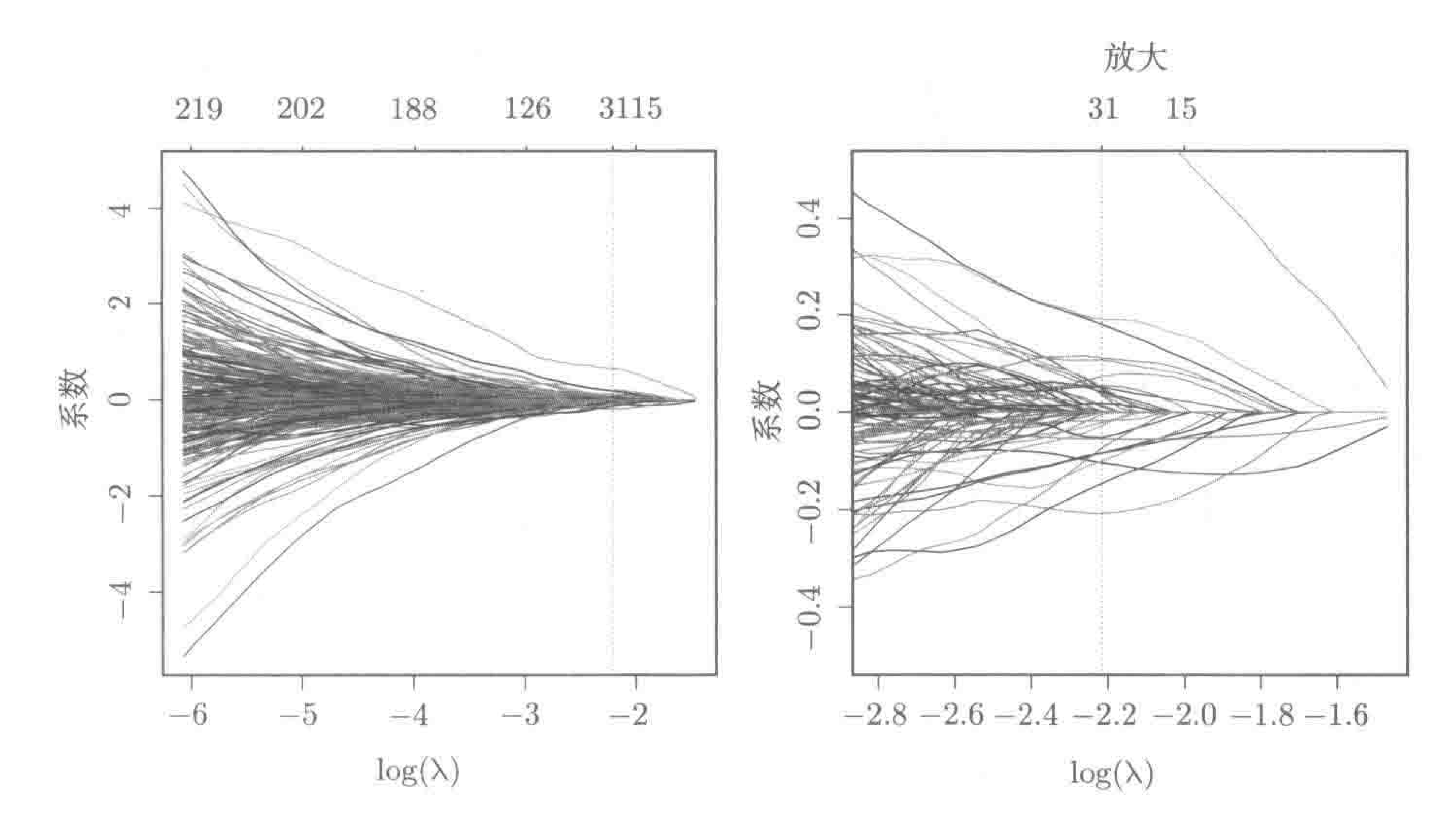


图 3-8 用 Cox 模型所拟合的淋巴瘤数据集的 ℓ_1 正则化系数路径。因为 $p \gg N$，所以图在靠近末端时呈喇叭形状，对应偏似然等于 1 的饱和模型。右图放大了重点区域，这是一种正则化程度很深的解，有 31 个非零系数

Cox 模型的求解与多分类模型相似，只是稍微复杂一点。Simon, Friedman, Hastie and Tibshirani（2011）给出了一种基于坐标下降法的详细求解方法。

3.5.1 交叉验证

本章的所有模型都需要考虑 λ 的选择，这里会采用 K 重交叉验证来选择 λ（如图 3-2 所示），其中 K 取 5 或者 10。对于 Cox 模型，需要计算出交叉验证的偏差，它等于负两倍的对数偏似然。如果 N/K 值小，计算偏差就会出现问题，即没有足够的样本可以计算风险集。这里会采用 van Houwelingen et al.（2006）给出的方法。在 k 已知时，可计算出系数 $\widehat{\beta}^{-k}(\lambda)$，然后计算

$$\widehat{\mathrm{Dev}}_{\lambda}^{k} := \mathrm{Dev}\left[\widehat{\beta}^{-k}(\lambda)\right] - \mathrm{Dev}^{-k}\left[\widehat{\beta}^{-k}(\lambda)\right] \tag{3.29}$$

右边第一项使用 N 个样本计算偏差，第二项去掉了 k 个样本。最终通过相减得到 $\mathrm{Dev}_{\lambda}^{CV} = \sum_{k=1}^{K} \widehat{\mathrm{Dev}}_{\lambda}^{k}$。这种方法的关键在于，每一项都有足够的样本来计算偏差，在标准情况下（即任何其他广义线性模型下），在剩下的数据集上的估计会得到精确的偏差。

图 3-9 中的偏差就是按照这种方式计算出来的。右图是放大后的结果。可以看出，在非零系数个数为 31 时，偏差取得最小值。图 3-7 展示了所选择模型的效果。对于每一个样本，要计算 $\widehat{\eta}(x_i) = x_i^{\mathrm{T}}\widehat{\beta}(\lambda_{\min})$，将这些评价分数按照阈值零分成两组。左图中两种颜色的生存曲线显示两组数据产生的生存曲线是一样的。它们被很

好地区分开来，也就是说这是一个有效的方法。但是这些评价分数有偏差：因为同样的数据既用来计算这些评价分数，又用来评估这些分数的性能。

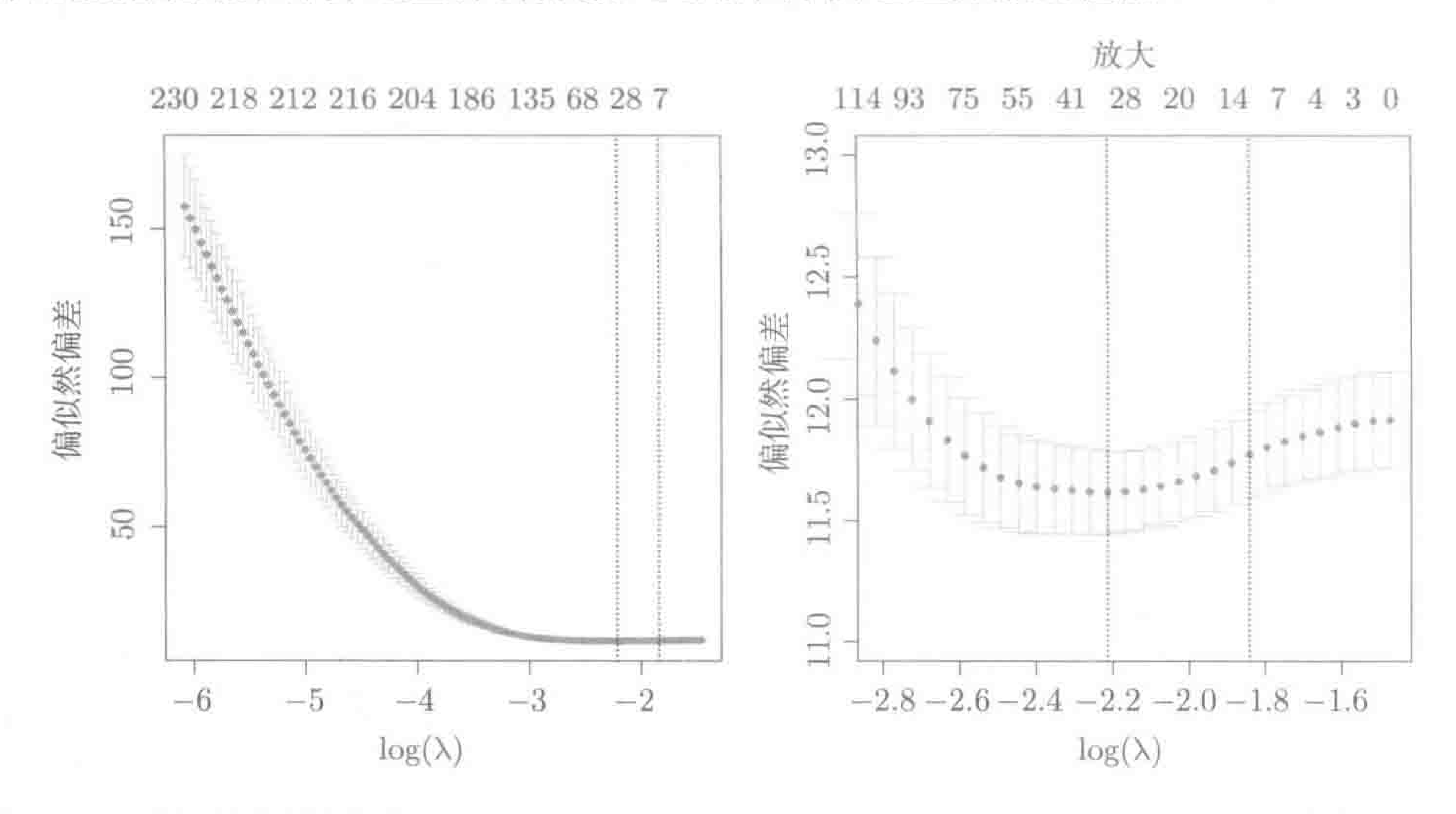


图 3-9　淋巴瘤数据上的交叉验证偏差，由文中介绍的相减方式计算得到。右图放大了重点区域。左图中的垂直虚线对应了最小值，这是本例所选的模型；右图对应了曲线最右边的点（最简单模型），它们与最小值在一个标准差内。以此为基础，可以选择较为保守的方法。图上方标出了非零参数的个数

3.5.2　预验证

在图 3-7 中，为了对模型做出公正的评估，采用了交叉验证的一个变种形式，即**预验证**（Tibshirani and Efron 2002）。交叉验证留出一部分数据，以便对模型的错误率进行合理的无偏估计。但是在有些情况下（如生存模型中），错误率并不是完全可解释的。预验证方法和交叉验证方法相似，但产生了一组新的“无偏数据集”，用于模拟模型在独立数据集上的表现。预验证数据集则可用于分析和展示。在计算 k 重样本的评价分数值 $\widehat{\eta}(x_i)^{(k)}$ 时，需要用那些被删除的样本 [1]来计算系数向量 $\widehat{\beta}^{(-k)}$。对所有的 K 重都这样做，就得到了“预验证”数据集 $\left\{\left(\widehat{\eta}(x_i)^{(k)}, y_i, \delta_i\right)\right\}_{i=1}^{N}$。预验证数据集的关键在于，评价分数 $\widehat{\eta}(x_i)^{(k)}$ 通过输出变量 (y_i, δ_i) 得到。因此，可以将这些分数看成是从一个完全独立的数据集得到的，这个数据集与“测试数据集”$\{(x_i, y_i, \delta_i)\}_{i=1}^{N}$ 全然不同。图 3-7 的右图将预验证分数分为两组，并画出了对应的生存曲线。尽管曲线并不像左图那样分得很开，但它们的区别依然很明显。

① 严格来说，每一次 λ 都应当重新选择，但这里我们并没有这样做。

3.6 支持向量机

下面介绍基于二分类的支持向量机（Support Vector Machine，SVM）。其算法思路如图 3-10 所示。决策边界是黄色区域内中间的那条实线。所谓**间隔**（margin）是指黄色区域的一半。理想情况下，所有的蓝点应当在黄色区域的右上方，红点应当在黄色区域的左下方。但图中有三个红点和两个蓝点在错误的间隔处，这就会得到相应的 "误差"ξ_i。SVM 的决策边界是由最大化间隔得到的，其总的误差 $\sum_{i=1}^N \xi_i$ 会限定在一个固定值内。SVM 的思想是：最大化间隔的决策边界能使两类之间有更多的空间，在测试数据上拥有更强的泛化能力。由此可以提出优化问题

$$\underset{\beta_0,\beta,\{\xi_i\}_1^N}{\text{maximize}} \quad M\text{，其约束为} \quad y_i \underbrace{(\beta_0 + \beta^{\mathrm{T}} x_i)}_{f(x_i;\beta_0,\beta)} \geqslant M(1-\xi_i) \forall_i \tag{3.30}$$

$$\xi_i \geqslant 0 \ \forall_i, \quad \sum_{i=1}^N \xi_i \leqslant C, \quad ||\beta||_2 = 1 \tag{3.31}$$

（对这种具体形式的解释见 3.6.1 节）。

这个问题是一个带凸约束的线性损失函数，可以使用很多高效算法求解。该问题等价于式（3.8）那样的惩罚形式，因此可重写为

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^N [1 - y_i f(x;\beta_0,\beta)]_+ + \lambda ||\beta||_2^2 \right\} \tag{3.32}$$

减小 λ 和减小 C 的效果一样 ①。线性 SVM 可以通过**核方法**得到非线性的决策边界；这需要用希尔伯特空间上的范数的平方来代替式（3.32）中的 ℓ_2 范数的平方，希尔伯特空间上的范数是由对称二元核定义的。这种推广可以在很多地方看到，例如 Hastie et al.（2009）Section 5.8。

因为式（3.32）中包含一个二次惩罚项，所以得到的系数向量不稀疏。但是，因为铰合损失函数为分段线性函数，这就得到了另一种形式的稀疏性。通过 SVM 的对偶变换可以得出解 $\widehat{\beta}$ 的形式为

$$\widehat{\beta} = \sum_{i=1}^N \widehat{\alpha}_i y_i x_i \tag{3.33}$$

每一个样本 $i \in \{1,\ldots,N\}$ 都有一个非负权重 $\widehat{\alpha}_i$，仅子集 $\mathcal{V}_\lambda$（即**支持向量集**）对应的权重都不为零。

① 式（3.32）的解并不满足 $||\widehat{\beta}||_2 = 1$，但线性分类器具有缩放不变性，所以系数可以再调整。

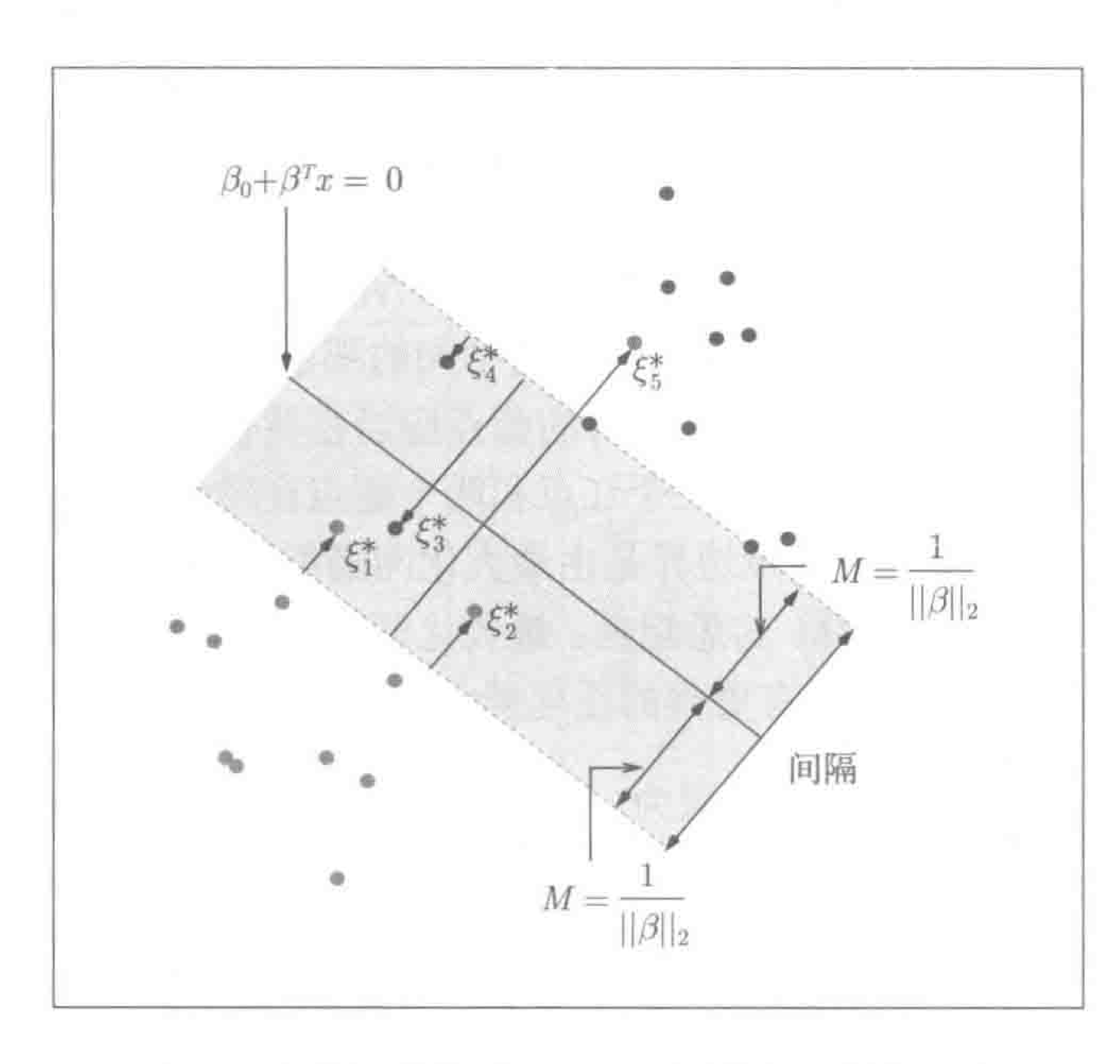


图 3-10　支持向量分类器：决策边界是实线，虚线围绕的区域为最大间隔，宽度为 $2M = 2/\|\beta\|_2$。标注为 ξ_j^* 的点表示误分类样本，其值 $\xi_j^* = M\xi_j$；正确分类样本的 $\xi_j^* = 0$。间隔的最大化受约束 $\sum_{i=1}^N \xi_i \leqslant C$ 的限制。因此 $\sum_{i=1}^N \xi_i^*$ 是所有误分类样本的距离总和

因为对线性和非线性核的计算复杂度都是 $\mathcal{O}(pN^2)$，所以 SVM 在高维分类问题（$p \gg N$）中非常流行。对于线性 SVM，用随机梯度法（Shalev-Shwartz, Singer and Srebro 2007）可以提高计算效率。但是该方法并无稀疏特征。将目标式（3.32）中的 ℓ_2 惩罚项代替为 ℓ_1 惩罚项，则为能提高稀疏性，这样就会得到 **ℓ_1正则化的线性 SVM**：

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} [1 - y_i f(x_i; \beta_0, \beta)]_+ + \lambda \|\beta\|_1 \right\} \tag{3.34}$$

式（3.34）是一个带有一些约束的线性优化问题（Zhu, Rosset, Hastie and Tibshirani 2004, Wang, Zhu and Zou 2006），其有效的求解算法较为复杂（见习题 3.9）。解的路径（细节上）有很多跳跃处，这源自其不连续性。因此，一些学者倾向于用处处可导的**平方铰合损失**函数 $\phi_{\text{sqh}}(t) = (1-t)_+^2$ 来替换传统的铰合损失函数 $\phi_{\text{hin}} = (1-t)_+$（见习题 3.8）。

SVM 的损失函数和二项式损失函数有很多相同点（Hastie et al. 2009, Section 12.3），它们的解并没有太大差别。图 3-11 用两个例子比较了二者的 ℓ_1 正则化的解的路径，由此可证明这一观点。在左图中，它们几乎重合。对于右图的大部分解，训练数据被解分离开来。在求解路径的最后面，SVM 系数要比逻辑斯蒂系数稳定

一些，二者在此处有较大差异。

另一方面，支持向量机用于寻找可分数据的最大间隔解，所以在求解路径的最后面，它的系数并不发散。但是，对于 ℓ_1 惩罚而言，求解路径的最后面会得到非稀疏解。因此，并不推荐用基于 ℓ_1 正则化的线性 SVM 来做特征选择，因为相应的逻辑斯蒂回归问题（3.6）在惩罚项起作用时也能得出非常相似的解，且算法更稳定。

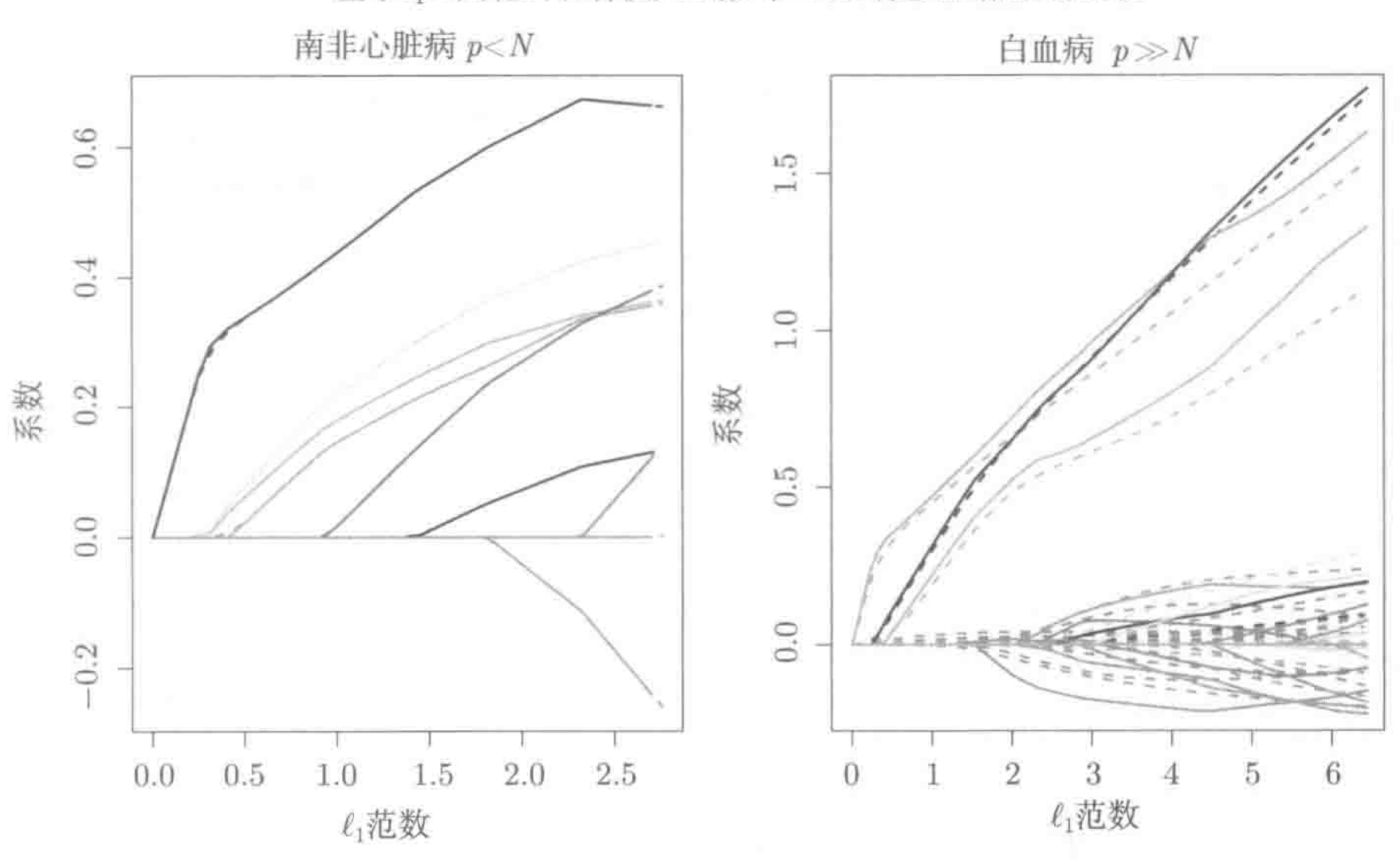


图 3-11 在两个不同数据集上比较基于 ℓ_1 正则化的 SVM 与逻辑斯蒂回归的系数路径。左图为南非心脏病数据（N=462，p=9），右图为白血病数据（N=38，p=6087）。虚线是 SVM 系数，实线是逻辑斯蒂回归系数。从两幅图中都可以看出它们的相似之处

可分数据与逻辑斯蒂回归

众所周知，在两类线性可分数据上，如果没有系数惩罚项，线性逻辑斯蒂回归模型会失败（见习题 3.1），因为系数的最大似然估计将无穷大。问题在于，似然函数试图将概率全置为 1 或者 0，而由式（3.2）可知，这在有限个参数下是不可能的。一旦加上惩罚项 [如式（3.6）]，则问题迎刃而解，因为只要 $\lambda > 0$，就不可能出现很大的系数。

宽数据（$N \ll p$）通常都可以让数据分类，除非在协变量空间上，两类有明显的联系。图 3-1 给出了宽数据情况下的逻辑斯蒂回归系数求解路径。注意，系数在最后面会呈扇形展开，因此需要注意系数路径最后面的情况，并且不能让 λ 取值太小。在很多情况下，这意味着糟糕的过拟合。本例中并没有出现这种情况，这种情况可在交叉验证示意中图看到（见图 3-2）。

由于二者都被当作最大间隔的分类器，因此在机器学习领域，求解路径的末尾有着特殊含义。在详细解释前，需先复习有关线性分类器的一些几何知识。线性分类器 $f(x) \equiv f(x;\beta_0,\beta) = \beta_0 + \beta^{\mathrm{T}}x$ 的决策边界为 $\mathcal{B} := \{x \in \mathbb{R}^p | f(x) = 0\}$。点 x_0 到决策边界的欧几里得距离为（见习题 3.2）

$$\mathrm{dist}_2(x_0, \mathcal{B}) := \inf_{z\in\mathcal{B}} ||z - x_0||_2 = \frac{|f(x_0)|}{||\beta||_2} \tag{3.35}$$

因此，给定一个样本和相应的输出 (x, y)，则 $\dfrac{yf(x)}{||\beta||_2}$ 为到决策边界的带符号的欧几里得距离，如果 y 和 $f(x)$ 的符号相反，则其值为负数。对于可分数据，最优分类超平面 $f^*(x) = 0$ 可通过求解优化问题

$$M_2^* = \max_{\beta_0,\beta} \left\{ \min_{i\in\{1,\dots,N\}} \frac{y_i f(x_i;\beta_0,\beta)}{||\beta||_2} \right\} \tag{3.36}$$

而得到。总而言之，要让最靠近边界的样本到边界的欧几里得距离最大。

最优分类超平面与存在一定限制的岭正则化逻辑斯蒂回归之间存在一种有趣的联系（Rosset et al. 2004）。实际上，假设将式（3.10）中的 ℓ_1 惩罚项替换为平方 ℓ_2 惩罚项，然后求解问题

$$\underset{\beta_0,\beta}{\mathrm{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \log(1 + \mathrm{e}^{-y_i f(x_i;\beta_0,\beta)}) + \lambda||\beta||_2^2 \right\} \tag{3.37}$$

令 $(\widetilde{\beta}_0(\lambda), \widetilde{\beta}(\lambda))$ 为最优解，则会得到一个特定的线性分类器。下面来看看随着正则化系数 λ 变小，线性分类器的变化情况。可以证明（Rosset et al. 2004）

$$\lim_{\lambda\to 0} \left\{ \min_{i\in\{1,\dots,N\}} \frac{y_i f(x_i;\widetilde{\beta}_0(\lambda),\widetilde{\beta}(\lambda))}{||\widetilde{\beta}(\lambda)||_2} \right\} = M_2^* \tag{3.38}$$

因此，基于 ℓ_2 正则化的逻辑斯蒂回归在求解路径最后面对应的就是 SVM 解。实际上，假设 $(\breve{\beta}_0, \breve{\beta})$ 为 SVM 目标函数式（3.30）在 C=0 时的解，则

$$\lim_{\lambda\to 0} \frac{\widetilde{\beta}(\lambda)}{||\widetilde{\beta}(\lambda)||_2} = \breve{\beta}. \tag{3.39}$$

ℓ_1 正则化模型会如何变化呢？问题有一点复杂，因为这里需要广义投影（general projection）和对偶范数（dual norm）的知识（Mangasarian 1999）。与式（3.35）中的 ℓ_2 距离一样，有如下的度量：

$$\mathrm{dist}_\infty(x_0, \mathcal{B}) := \inf_{z\in\mathcal{B}} ||z - x_0||_\infty = \frac{|f(x_0)|}{||\beta||_1} \tag{3.40}$$

对于给定的 $\lambda \geqslant 0$，设 $(\widehat{\beta}_0(\lambda), \widehat{\beta}(\lambda))$ 为 ℓ_1 正则化逻辑斯蒂回归目标函数（3.10）的最优解。当 λ 趋向于零时，有

$$\lim_{\lambda \to 0} \left\{ \min_{i \in \{1,\ldots,N\}} \frac{y_i f(x_i; \widehat{\beta}_0(\lambda), \widehat{\beta}(\lambda))}{||\widehat{\beta}(\lambda)||_1} \right\} = M_\infty^* \tag{3.41}$$

因此在最坏的情况下，ℓ_1 正则化逻辑斯蒂回归模型的间隔将收敛到基于 ℓ_1 正则化的支持向量机，该模型将 ℓ_∞ 间隔 [见式（3.40）] 最大化。

总之，可以得到以下结论。

- 在求解路径的最后面，即解的最稠密处，逻辑斯蒂回归的解与 SVM 解一致。
- 在该区域，有更稳定的数值算法可以求解 SVM 问题。
- 相比之下，在求解路径较稀疏的部分，逻辑斯蒂回归最实用。

3.7 计算细节及 glmnet

本章中的大多数示例均使用 R 语言包 glmnet（Friedman et al. 2015）求解得到。下面将详细介绍 glmnet 包的一些选项及性质。尽管这些都是该包所特有的，但其他类似软件也自然需要这些选项和性质。

family：family 选项用于选择损失函数和相关模型。1.7 版有 gaussian、binomial、multinomial （grouped or not）、poisson 及 cox。gaussian 族允许多个响应（多任务学习），在这种情况下，组 lasso 像在组多项式回归（grouped multinomial）中一样，用于为每一个变量选择系数。deviance 统计量与每个族相关，类似于高斯误差中的残差平方和。$\widehat{\boldsymbol{\mu}}_\lambda$ 表示参数为 λ 时拟合后的 N 维均值向量，$\widehat{\boldsymbol{\mu}}$ 为无约束或者饱和拟合模型，则有

$$\text{Dev}_\lambda \doteq 2[\ell(\boldsymbol{y}, \widetilde{\boldsymbol{\mu}}) - \ell(\boldsymbol{y}, \widetilde{\boldsymbol{\mu}}_\lambda)] \tag{3.42}$$

这里 $\ell(\boldsymbol{y}, \boldsymbol{\mu})$ 是指模型 $\boldsymbol{\mu}$ 的对数似然，是 N 项的和。**空模型偏差**（null deviance）是指 $\text{Dev}_{\text{null}} = \text{Dev}_\infty$，通常意味着均值 $\widehat{\boldsymbol{\mu}}_\infty = \bar{y}\mathbf{1}$，cox 族在这种情况下的 $\widehat{\boldsymbol{\mu}}_\infty = \mathbf{0}$。glmnet 会得到 D^2，指可解释偏差部分，见式（3.11）定义。

惩罚：对于所有的模型，glmnet 算法允许一定范围内的弹性网惩罚，惩罚范围在 ℓ_1 和 ℓ_2 之间。惩罚优化问题的广义形式为

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ -\frac{1}{N}\ell(\boldsymbol{y}; \beta_0, \beta) + \lambda \sum_{j=1}^{p} \gamma_j \{(1-\alpha)\beta_j^2 + \alpha|\beta_j|\} \right\} \tag{3.43}$$

这一系列的惩罚项主要由如下三组参数决定。

- 参数 λ 决定了模型的整体复杂度。默认情况下，glmnet 算法会产生 100 个 λ 值序列，覆盖全部求解解径（在 log 尺度下），需注意饱和模型的开始处。

- 弹性网参数 $\alpha \in [0,1]$，将岭回归与 lasso 回归组合在一起。尽管可以通过交叉验证来选择 α 参数，但通常取 3~5 个 α 值。
- 对于 $j = 1,2,\ldots,p, \gamma_j \geqslant 0$ 是一个惩罚调节量。当 $\gamma_j = 0$ 时，第 j 个变量总是包含在模型中，当 $\gamma_j = \inf$ 时，则该变量总是被排除。通常情况下 $\gamma_j = 1$（默认），即对所有变量做相同处理。

系数边界：使用坐标下降法可以很便捷地求得模型中各个参数的上界和下界。例如，需要 lasso 模型非负。在这种情况下，以此简单设定边界，可以解决坐标循环过程中有参数超出上界或下界的问题。

偏移量：所有模型都允许存在偏移量。对于每一个样本来而言，这是一个实数值 o_i，要加到线性特征中。这与其他参数无关：

$$\eta(x_i) = o_i + \beta_0 + \beta^{\mathrm{T}} x_i \tag{3.44}$$

偏移量有很多用处。有时我们需要查看已经拟合好的模型 $h(z)$（z 中可能包含 x 或者与 x 相一致），想看加入一个线性模型是否会提升效果。这时就可以为每个样本提供一个 $o_i = h(z_i)$。

对于泊松模型，如果对每个样本的观察周期不同，则引入偏移量可建立比率模型，而非对计数均值建模。假设在 t 时间内观察到计数值 Y，则 $\mathbb{E}[Y|T = t, X = x] = t\mu(x)$。其中 $\mu(x)$ 是单位时间内的比率。使用 log 联接，可为每个样本提供 $o_i = \log(t_i)$。具体例子见 3.4.1 节。

矩阵输入和权重：二项式和多项式模型响应变量通常以 2 或者 K 水平因子的形式给出。另外，glmnet 允许响应变量以矩阵形式提供。这主要针对归组数据，在这种情况下，每个 x_i 均为一个多项式样本。$N \times K$ 响应矩阵每行中的值代表属于该类的数量。另外，这些行总和为 1。对于后一种情况，若提供样本权重等于每个样本总数，则等价于第一种形式。在**未分组**的形式下，一个指示响应矩阵等同于将数据视为因子。

稀疏模型矩阵 $\boldsymbol{X}$：通常，当 $p \gg N$ 时，输入矩阵 $\boldsymbol{X}$ 中会存在很多零。例如，在文本模型中，每个特征向量 $x_i \in \mathbb{R}^p$ 是每个词在文本中出现的次数，这是从一部很大的词典中统计得到的。对这种向量或矩阵的有效存储方式是只保存非负值，这样由行和列值就可以索引到对应值。坐标下降法非常适合利用这种稀疏性，因为可以一次处理一个变量，并且基本操作是计算内积。例如，在 3.4.1 节中，模型矩阵 $\boldsymbol{X} = \boldsymbol{I}$ 是一个非常稀疏的单位矩阵，其大小为 $N \times N$。即使 $N = 10^6$，程序依然能够在 27 秒内计算出 100 个松弛值 δ 的结果。

参考文献注释

广义线性模型由 Nelder and Wedderburn（1972）作为一类模型提出，McCullagh

and Nelder（1989）给出了全面解释。Tibshirani （1996）提出将 lasso 方法用于逻辑斯蒂回归。Friedman, Hastie, Hoeflingt and Tibshirani（2007）、Friedman, Hastie and Tibshirani（2010b）、Wu and Lange（2008），以及 Wu, Chen, Hastie, Sobel and Lange（2009）将坐标下降法用于逻辑斯蒂回归、多项式回归以及泊松回归。Tibshirani and Efron（2002）提出了预验证方法。Boser, Guyon and Vapnik（1992）提出了支持向量机思想，Vapnik（1996）进行了完全实现。

习　题

习题 3.1　在一个可分数据（即数据能够由超平面正确地分为两类）上，如何采用线性逻辑斯蒂回归进行分类？求证：似然估计无界，且对数似然目标函数会得到的最大值为零。在这种情形下，拟合概率值是否具有意义？

习题 3.2　响应变量 $y \in \{-1, +1\}$，线性分类器函数为 $f(x) = \beta_0 + \beta^{\mathrm{T}} x$，假设根据 $\operatorname{sgn}(f(x))$ 来分类。求证：点 x 与类标签 y 到决策界之间的带符号的欧几里得距离为

$$\frac{1}{||\beta||_2} y f(x) \tag{3.45}$$

习题 3.3　对于多项式模型，求证：惩罚项会自动对参数估计进行归一化。这里需要求解基于广义弹性网惩罚的问题（见 4.2 节）。对于某个参数 $\alpha \in [0,1]$，考虑

$$c_j(\alpha) = \underset{t \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{\ell=1}^{K} \left[\frac{1}{2}(1-\alpha)(\beta_{j\ell} - t)^2 + \alpha|\beta_{j\ell} - t| \right] \right\} \tag{3.46}$$

设 $\overline{\beta}_j = \frac{1}{K}\sum_{l=1}^{K} \beta_{j\ell}$ 为样本均值，$\widetilde{\beta}_j$ 为样本中值。（简单起见，假设 $\overline{\beta}_j \leqslant \widetilde{\beta}_j$）。求证：

$$\overline{\beta}_j \leqslant c_j(\alpha) \leqslant \widetilde{\beta}_j, \quad \alpha \in [0,1] \tag{3.47}$$

当 α=0 时，左边不等式成立；当 α=1 时，右边不等式成立。

习题 3.4　请推导**最大熵**问题（3.24）中的拉格朗日对偶（3.25）。注意，因为在目标函数式（3.24）中，对数函数是一个限制，所以它自然为正。（提示：可以引入变量 $w_i = p_i - r_i$，这样在约束条件 $w_i = p_i - r_i$ 下，通过 $\{p_i, w_i\}_{i=1}^{N}$ 最小化式（3.24）。）

习题 3.5　请思考最大熵问题中的对偶问题（3.25）及其相关例子。假设对每一个单元可计算对应连续域 x 的中值单元的值 x_k。请考虑模型

$$q_k = u_k \mathrm{e}^{\beta_0 + \sum_{m=1}^{M} \beta_m x_k^m + \alpha_k} \tag{3.48}$$

假设用式（3.25）没有任何系数的惩罚对数似然来拟合。求证：对于估计分布 $\widehat{\boldsymbol{q}} = \{\widehat{q}_k\}_{k=1}^K, X$ 的矩为 M 阶时，与经验分布 $\boldsymbol{r} = \{r_k\}_{k=1}^N$ 匹配。

习题 3.6　对基于组 lasso 正则化的多项式回归 [见式（3.20）]，假设对于某一特定 λ，系数 $\widehat{\beta}_{kj}$**不等于**0。求证：对所有的 $l \in (1,\ldots,K), \widehat{\beta}_{lj} \neq 0$，且有 $\sum_{l=1}^K \widehat{\beta}_{lj} = 0$。

习题 3.7　该习题与组 lasso 正则化多项式回归 [见式（3.20）] 相关。假设对于某一特定 λ，拟合概率值为 $\widehat{\pi}_i = (\widehat{\pi}_{i1},\ldots,\widehat{\pi}_{iK})^{\mathrm{T}}$。设 $r_i = (r_{i1},\ldots,r_{iK})^{\mathrm{T}}$ 为观察到的比例值。现有一变量（向量）Z，观察值是 z_i，并希望更新这个拟和。设 $g = \sum_{i=1}^N z_i(r_i - \widehat{\pi}_i)$。求证：如果 $||g||_2 < \lambda$，则 Z 的系数为零，模型不变。

习题 3.8　对于二分类问题，**平方铰合损失函数**$\phi_{\mathrm{sqh}}(t) := (1-t)_+^2$ 可作为间隔损失函数 $\phi(yf(x))$ 的基函数。

(a) 求证：ϕ_{sqh} 处处可微。

(b) 假设 $Y \in \{-1,+1\}, \Pr(Y=1) = \pi \in (0,1)$。找出函数 $f:\mathbb{R}^p \to \mathbb{R}$，使其让下面的目标函数最小化（$x \in \mathbb{R}^p$ 成立）

$$\underset{f}{\text{minimize}}\, \mathbb{E}_Y[\phi_{\mathrm{sqh}}(Yf(x))] \tag{3.49}$$

(c) 用常见铰合损失函数 $\phi_{\mathrm{hin}}(t) = (1-t)_+$ 重复证明 (b)。

习题 3.9　给定二值响应变量 $y_i \in \{-1,+1\}$，考虑 ℓ_1 正则化的 SVM 问题：

$$(\widehat{\beta}_0, \widehat{\beta}) = \underset{\beta_0,\beta}{\arg\min}\left\{\sum_{j=1}^N \{1 - y_i f(x_i;\beta_0,\beta)\}_+ + \lambda \sum_{j=1}^p |\beta_j|\right\} \tag{3.50}$$

其中 $f(x_i;\beta_0,\beta) := \beta_0 + \beta^{\mathrm{T}}x$。比较该问题的解与加权 ℓ_2 正则化的 SVM 问题的解。给定非负权值 $\{w_j\}_{j=1}^p$，求解

$$(\widetilde{\beta}_0, \widetilde{\beta}) = \underset{\beta_0,\beta}{\arg\min}\left\{\sum_{i=1}^N \{1 - y_i f(x_i;\beta_0,\beta)\}_+ + \frac{\lambda}{2} \sum_{j=1}^p w_j \beta_j^2\right\} \tag{3.51}$$

(a) 求证：如果式（3.51）中 $w_j = 1/|\widehat{\beta}_j|$，则 $(\widetilde{\beta}_0, \widetilde{\beta}) = (\widehat{\beta}_0, \widehat{\beta})$。

(b) 对于给定的权值序列 $\{w_j\}_{j=1}^p$，对于所有的 $j = 1,\ldots,p, \omega_j \in (0,\infty)$，如何用普通未加权的 SVM 去求解式（3.51）？如果对一些子集 $w_j = \infty$，又该如何求解？

(c) 借鉴前面的部分，用普通的 SVM 设计求解式（3.50）的迭代算法。

第 4 章　广义 lasso 惩罚

4.1　引　　言

第 3 章通过各种损失函数介绍了 lasso 惩罚的推广形式。本章将介绍一些基于 lasso ℓ_1 惩罚项本身的实用变体，扩大基础模型的使用范围。这些变体均继承了标准 lasso 的两个基本特征：收缩性（shrinkage）和变量（组）选择。

这种广义惩罚可应用于很多情况。微阵列分析中便经常有相关特征组，比如作用于同一生物通路（biological pathway）的基因。实验表明，在高度相关的特征中，lasso 有时表现不佳。将平方 ℓ_2 惩罚项和 ℓ_1 惩罚项结合在一起，就可得到**弹性网**惩罚，这种惩罚能够更好地处理相关特征组，且会选择全部相关特征或不选任何相关特征。在另一些应用中，可以按结构对特征进行分组，例如哑（dummy）变量，用于对多层的分类特征或者多元回归问题中的一组系数进行编码。这种情况自然要全部选择（或忽略）一组系数。**组 lasso** 和 **重叠组 lasso**（overlap group lasso）通过将（非平方）ℓ_2 惩罚项相加做到这一点。另一种结构组来自潜在的索引，比如时间。每个参数都可能有一个相关的时间戳，因此可以认为时间相邻系数可能相同或者相似。**融合 lasso**（fused lasso）就是为这种问题而设计的。

最后，有很多非参数光滑方法能处理大的分组变量。例如，一个相加的光滑样条模型中的每一项都有一个相关的三次样条基。组 lasso 方法能自然地应用于这些情况，COSSO 和 SPAM 族就属于这类非参数模型。总之，这些变体都能自然地处理不同情况下的分组特征，本章要详细介绍它们。

4.2　弹性网惩罚

lasso 方法不擅长处理高度相关变量，系数路径会不稳定，有时会出现一些奇怪的行为。考虑一个简单而典型的例子，变量 X_j 在某个 λ 下拟合的模型系数为 $\widehat{\beta}_j > 0$。如果在原数据上增加 $X_{j'} = X_j$，则它们之间可用无数种组合方式共享系数：$\widetilde{\beta}_j + \widetilde{\beta}_{j'} = \widehat{\beta}_j$，这两部分都为正，这与损失函数和 ℓ_1 惩罚无关。因此，这一对变量的系数不确定。而二次惩罚项则会精确地将 $\widehat{\beta}_j$ 平分给这两个变量（见习题 4.1）。事实上，不太可能有一对完全相同的变量，但是通常会有非常相关的变量组。在微阵列分析中，同一基因通路中的基因组趋向于共同起作用，因此这些表达式的测量

极为相关。图 4-1 的左边是这种情况下的 lasso 系数路径图。数据分为两组，每组 3 个变量，组内变量之间的相关系数为 0.97。样本数为 N=100，数据通过下面公式模拟得到

$$
\begin{aligned}
&Z_1, Z_2 \sim N(0,1)\text{独立} \\
&Y = 3Z_1 - 1.5Z_2 + 2\varepsilon, \quad \varepsilon \sim N(0,1) \\
&X_j = Z_1 + \xi_j/5, \xi_j \sim N(0,1) \quad j = 1,2,3 \\
&X_j = Z_2 + \xi_j/5, \xi_j \sim N(0,1) \quad j = 4,5,6
\end{aligned} \tag{4.1}
$$

如图 4-1 左图所示，lasso 估计并不能反映变量的重要性。

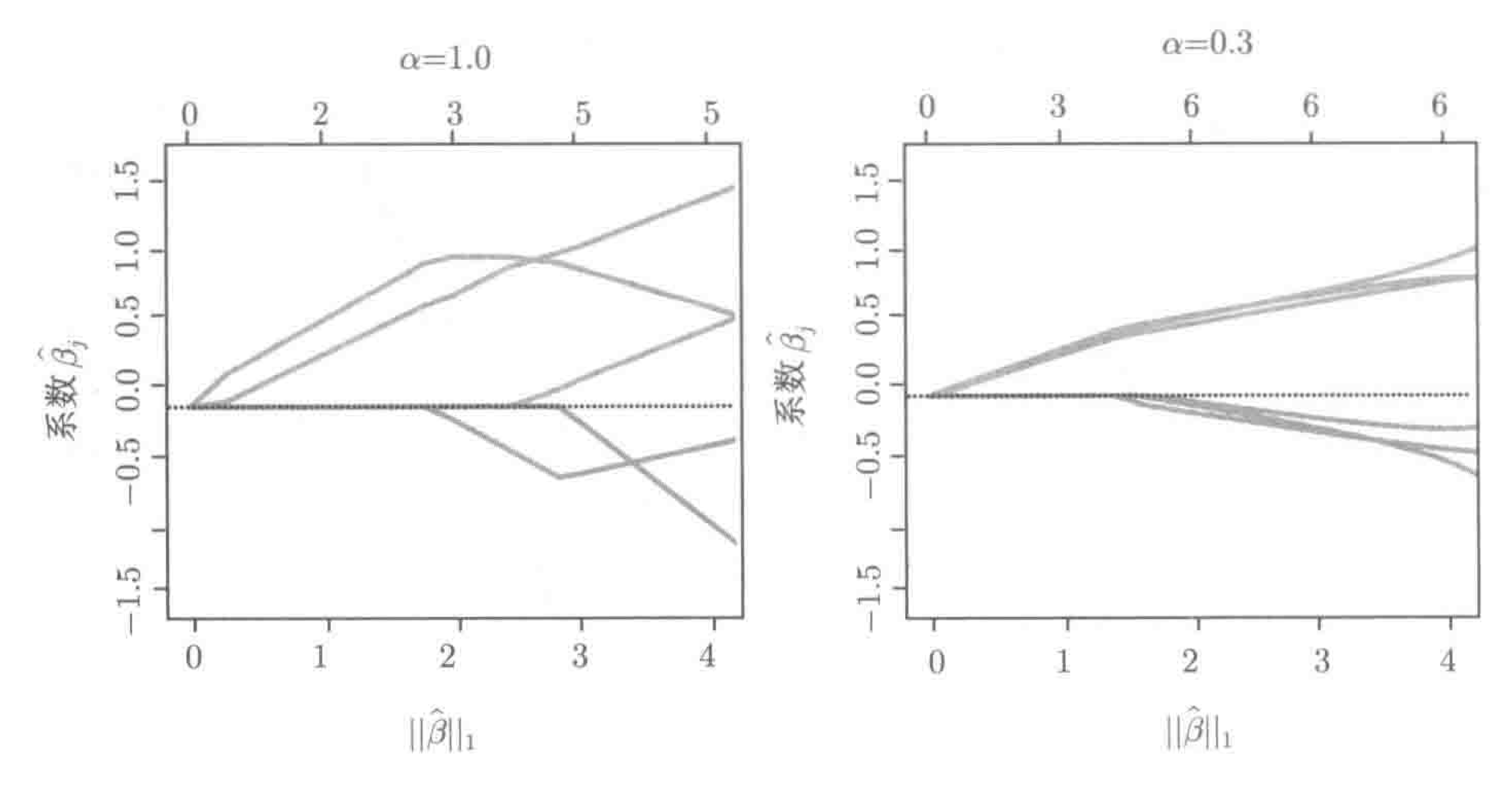


图 4-1　6 个变量，3 个一组，同组变量高度相关。左图为 lasso 估计（α=1），随着正则化参数 λ 的变化，可看出这些估计并不具有规律。右图为弹性网估计（α=0.3），包含所有变量，相关的组被归到一起

弹性网是岭惩罚和 lasso 惩罚之间的折中（Zou and Hastie 2005）；它要求解下面的凸优化问题：

$$
\underset{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{minmize}} \left\{ \frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^{\mathrm{T}}\beta)^2 + \lambda\left[\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1\right] \right\} \tag{4.2}
$$

其中 $\alpha \in [0,1]$ 是一个可变参数。单个参数上的弹性网惩罚为（忽略正则化参数 $\lambda>0$）

$$
\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \tag{4.3}
$$

当 α=1 时，惩罚项为 ℓ_1 正则化（也称 lasso 惩罚）；当 α=0 时，为平方 ℓ_2 正则化（也称岭惩罚）；[①]

① 弹性网惩罚二次部分中的 $\frac{1}{2}$ 在优化中是一个更加直观的软阈值因子。

回到图 4-1，右图是 α=0.3 时弹性网惩罚下的系数路径图。与左图中基于 lasso 的系数路径相比，同组系数相近，其值也相近。当然，这个例子是理想化的，事实上组的结构不会这么清晰。但是，通过在 ℓ_1 惩罚项上加上岭惩罚，弹性网惩罚会自动控制每组变量间的强关联。另外，对于任意的 $\alpha < 1$ 和 $\lambda > 0$，弹性网问题（4-2）是严格凸的：不管 X_j 是否相关都存在唯一解。

图 4-2 比较了弹性网（左图）和 lasso（右图）在 3 个变量时的约束区域。从图中可看出，弹性网球同时拥有 ℓ_2 球和 ℓ_1 球的特征：尖角和边会进行选择，弯曲的轮廓会共享系数。习题 4-2 会进一步研究这些性质。

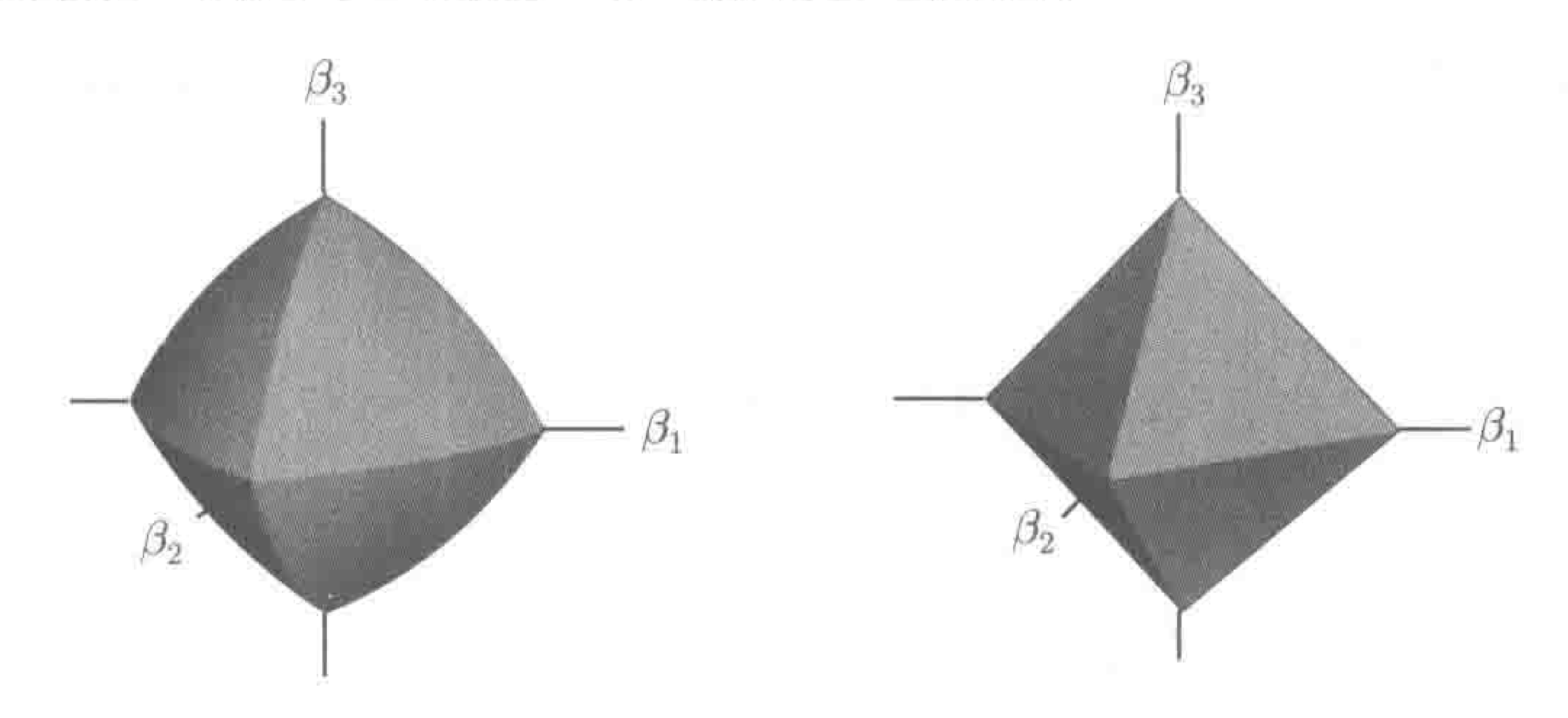


图 4-2 左图为 $\mathbb{R}^3$ 中 $\alpha = 0.7$ 时的弹性网球，右图为 ℓ_1 球。球上弯曲的轮廓迫使强相关的变量共享系数（详见习题 4.2）

弹性网方法还需要确定一个附加的调整参数 α。事实上，这可以看作是一个更高级的参数，可以主观设定。另外，交叉验证可以包括 α 值的一个（粗）网格。

式（4.2）中凸的弹性网问题基于参数对 $(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p$，有许多算法可以用来求解这个问题。坐标下降法尤其有效，其迭代过程是对 lasso 方法（见第 2 章）的简单扩展。模型中含有不需惩罚的截距，在开始阶段可以忽略，对协变量 x_{ij} 进行简单的归一化处理，然后得到最优的截距为 $\widehat{\beta}_0 = \bar{y} = \frac{1}{N}\sum_{j=1}^N y_j$。求解出最优截距 $\widehat{\beta}_0$ 后，需要计算最优向量 $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$。可以验证（习题 4.3）坐标下降法的第 j 次迭代的系数为

$$\widehat{\beta}_j = \frac{\mathcal{S}_{\lambda\alpha}\left(\sum_{i=1}^N r_{ij}x_{ij}\right)}{\sum_{i=1}^N x_{ij}^2 + \lambda(1-\alpha)} \tag{4.4}$$

其中 $\mathcal{S}_{\lambda\alpha} := \text{sgn}(z)(z-u)_+$ 是软阈值因子，$r_{ij} := y_i - \widehat{\beta}_0 - \sum_{k\neq j} x_{ij}\widehat{\beta}_k$ 是部分残差。我们循环迭代式（4.4）直到收敛。Friedman et al.（2015）给出了更多细节，提供了对不同损失函数的弹性网惩罚的具体实现方法。

4.3　组 lasso

在很多回归问题中，协变量本身就具有组结构。人们希望组内的所有参数同时不为零，或者同时为零，因此设计了不同形式的组 lasso 惩罚来实现这一目标。一个典型的例子就是当预测子中有定性因子时，就会出现组结构。通常用一组哑变量或者对照变量来对水平因子编码，让模型同时包含或者排除这一组变量。下面先来定义组 lasso，然后再扩展到其他例子中。

考虑一个线性回归模型，包含 J 组协变量，其中 $j=1,\ldots,J$，向量 $Z_j \in \mathbb{R}^{p_j}$ 表示第 j 组协变量。这里的目标是基于协变量集合 $(Z_1,\ldots,Z_J)$ 预测实数响应变量 $Y \in \mathbb{R}$。线性回归模型 $\mathbb{E}(Y|Z)$ 的形式为 $\theta_0 + \sum_{j=1}^{J} Z_j^{\mathrm{T}}\theta_j$，其中 $\theta_j \in \mathbb{R}^{p_j}$ 表示一组回归系数 p_j。①

给定一组样本数为 N 的样本集 $\{(y_i, z_{i1}, z_{i2}, \ldots, z_{iJ})\}_{i=1}^{N}$，组 lasso 需要求解下面的凸优化问题：

$$\underset{\theta_0\in\mathbb{R},\theta_j\in\mathbb{R}^{p_j}}{\text{minmize}} \left\{ \frac{1}{2}\sum_{i=1}^{N}\left(y_i-\theta_0-\sum_{j=1}^{J} z_{ij}^{\mathrm{T}}\theta_j\right)^2 + \lambda\sum_{j=1}^{J}\|\theta_j\|_2 \right\} \tag{4.5}$$

其中 $\|\theta_j\|_2$ 是向量 θ_j 的欧几里得范数。

广义组 lasso 有如下性质。

- 对于 $\lambda \geqslant$ 零，$\widehat{\theta}_j$ 向量或者整个为零，或者全不为零；②
- 当 $p_j=1$ 时，有 $\|\theta_j\|_2 = |\theta_j|$，所以如果所有组都只有一个变量，优化问题（4.5）就会变为普通的 lasso 问题。

图 4-3 基于 3 个变量比较组 lasso（左图）与 lasso（右图）的约束区域。从该图可以看出，组 lasso 球有 ℓ_2 球和 ℓ_1 球的特征。

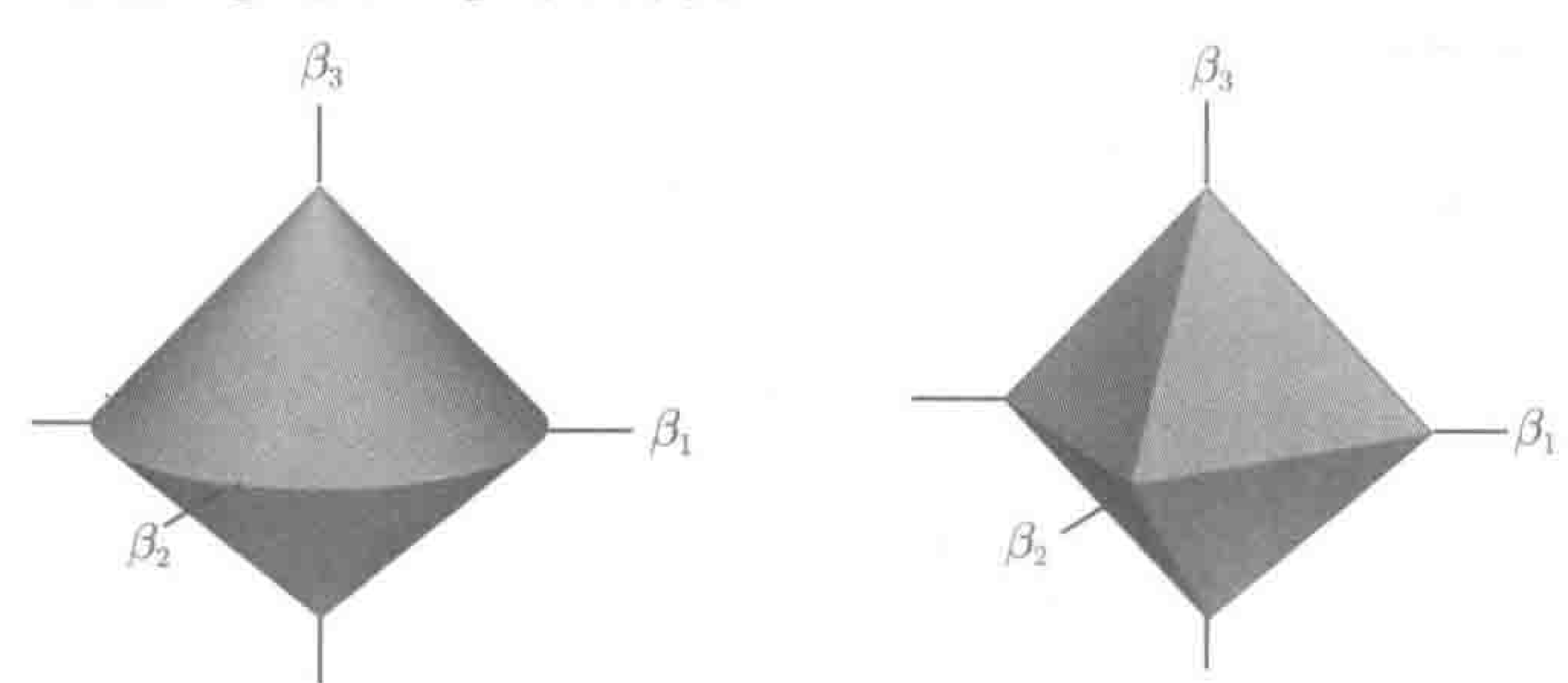


图 4-3　$\mathbb{R}^3$ 中的组 lasso 球（左图）和 ℓ_1 球（右图）。这里有两个组，系数分别为 $\theta_1 = (\beta_1,\beta_2) \in \mathbb{R}^2$，$\theta_2 = \beta_3 \in \mathbb{R}^1$

① 为了避免混淆，我们用 Z_j 和 θ_j 表示组变量及其系数，而 X_j 和 β_j 则用来表示标量。

② 一般全不为零，尽管特殊的结构可能会导致一组内的某些系数为 0，就像线性回归或岭回归中一样。

在式（4.5）中，所有组都受到相同的惩罚，有一种方法可以让较大的组更有可能被选择到。Yuan and Lin（2006a）最早建议根据各组的大小，由因子 $\sqrt{p_j}$ 来对各组进行加权惩罚。在相关例子中，组矩阵 $\boldsymbol{Z}_j$ 是标准正交的，对于一般矩阵，人们对因子 $\|\boldsymbol{Z}_j\|_F$ 进行讨论（见习题 4.5）。这些选择有些主观，但比较容易调整。简单起见，本文不讨论这方面的调整方法。

下面用几个例子来解释组 lasso（4.5）的应用。

例 4.1：多级因子回归 当线性回归中的变量是多级因子时，通常会为因子的各级加上单独的系数。举个简单的例子，有一个连续变量 X 和三级因子 G，其各级分别为 g_1, g_2, g_3。均值的线性模型为

$$\mathbb{E}(Y\,|X,G) = X\beta + \sum_{k=1}^{3} \theta_k \mathbb{I}_k[G] \tag{4.6}$$

其中 $\mathbb{I}_k[G]$ 是取值为 0 或 1 的指示函数，用来表示事件 $\{G = g_k\}$ 值。模型（4.6）为线性回归，变量为 X，截距 θ_k 依赖于 G 的不同级。

引入含有三个**哑变量**的向量 $Z = (Z_1, Z_2, Z_3)$，其中 $Z_k = \mathbb{I}_k[G]$，就可通过一个标准线性回归将这个模型表示为

$$\mathbb{E}(Y\,|X,G) = \mathbb{E}(Y\,|X,Z) = X\beta + Z^{\mathrm{T}}\theta \tag{4.7}$$

其中 $\theta = (\theta_1, \theta_2, \theta_3)$。在这里，$Z$ 是一个组变量，表示单因子 G。如果变量 G（通过向量 Z 来编码）没有任何预测作用，则整个向量 $\theta = (\theta_1, \theta_2, \theta_3)$ 全为零。另一方面，如果 G 对预测值有用，一般会期望向量 θ 的系数全不为零。对于更常见的情形，会有许多这种单变量或者组变量，所以模型形式为

$$\mathbb{E}(Y\,|X,G_1,\ldots,G_J) = \beta_0 + X^{\mathrm{T}}\beta + \sum_{j=1}^{J} Z_j^{\mathrm{T}}\theta_j \tag{4.8}$$

当为这种模型选择变量时，通常要按组包括或者排除组变量而非单独变量，组 lasso 模型的设计正符合这样的要求。

对于没有带惩罚的因子线性回归，需要注意可能出现的混淆。在这个例子中，一组哑变量会加 1，这会与截距混淆。可以用对照变量来对因子编码，例如，迫使一组内的系数和为零。因为带有 ℓ_2 惩罚项，所以这对组 lasso 来说不是问题。由于惩罚项确保了一个组内的全部系数之和为零，因而可用上面那种对称全表示法（见习题 4.4）。

变量也有可能按其他原因分组。例如，在基因表达阵列中，可能会有一组来自同一基因通路的高相关基因。选择这一组就相当于选择了这个基因通路。图 4-4 展示了基于基因数据的组 lasso 拟合的系数路径，这种拟合主要用来进行剪接位点（splice-site）检测（Meier, van de Geer and Bühlmann 2008，Section 5）。数据

来自人类 DNA，每个样本包含 7 组变量，值为 $\{A,G,C,T\}^7$。一些样本位于外显（exon-intron）子区域（剪接位点），另一些则不是，可用二值响应进行编码。Burge and Karlin（1977）对数据集进行了详细介绍。该回归问题要利用 7 个四因子变量 G_j 来预测二值响应变量 Y，每一类包含 5610 个训练样本。

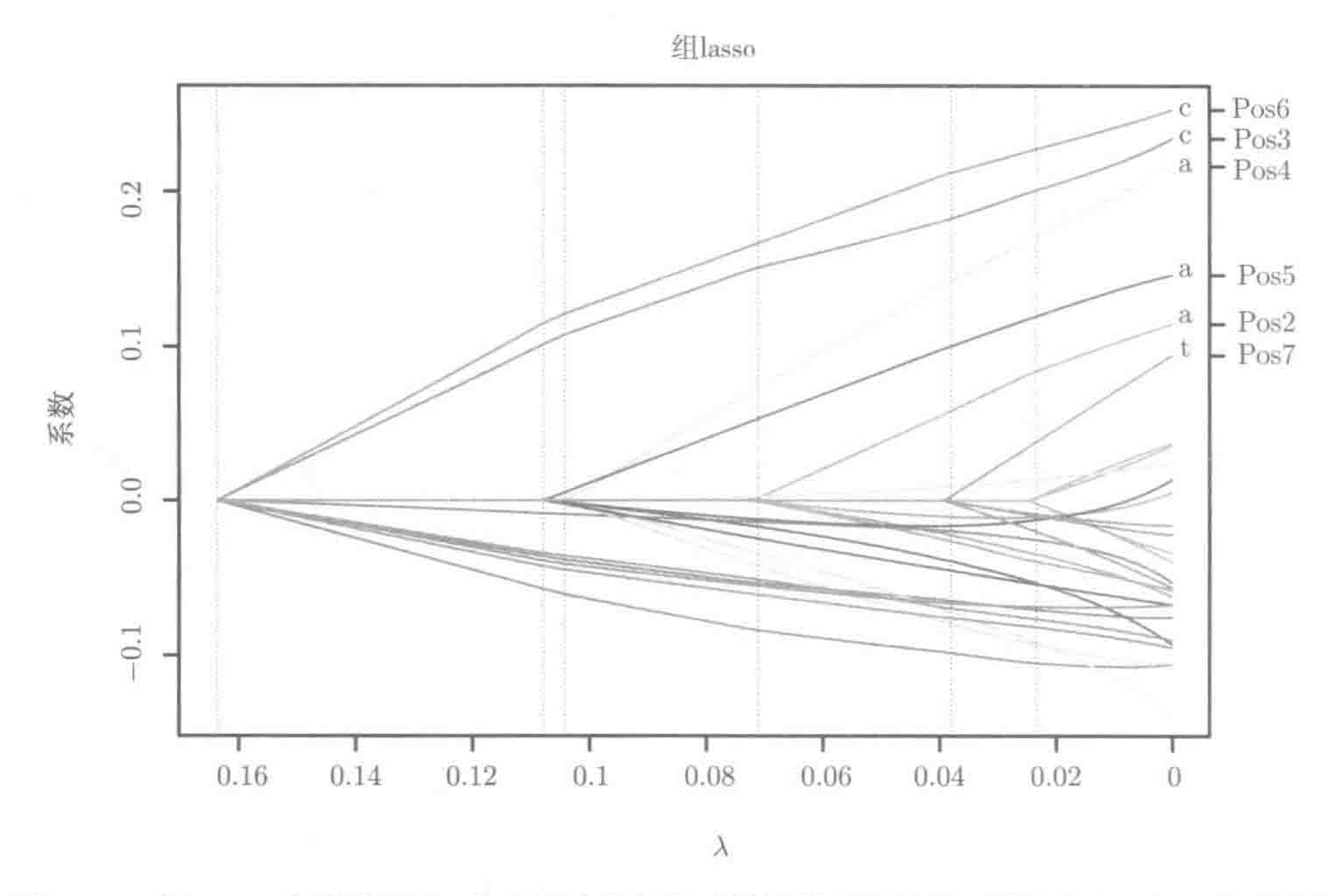


图 4-4 组 lasso 的系数路径，这是拟合剪接位点检测数据的结果。系数分为四组，分别对应核苷酸 A、G、C、T。垂直线表示一组进入。右边标注了一些变量，例如 Pos6 和级别 c。同一组中的系数颜色相同，平均值总是为 0

例 4.2：多元回归 有时需要基于一组预测子向量 $\boldsymbol{X} \in \mathbb{R}^p$ 来预测多变量响应 $\boldsymbol{Y} \in \mathbb{R}^K$（即**多任务学习**）。对给定 N 个样本 $\{(y_i, x_i)\}_{i=1}^N$，设矩阵 $\boldsymbol{Y} \in \mathbb{R}^{N\times K}$，$\boldsymbol{X} \in \mathbb{R}^{N\times p}$，这两个矩阵的第 i 行分别用 y_i 和 x_i 表示。如果用线性模型来拟合数据集，则可写为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{E} \tag{4.9}$$

其中 $\boldsymbol{\Theta} \in \mathbb{R}^{p\times K}$ 是系数矩阵，$\boldsymbol{E} \in \mathbb{R}^{N\times K}$ 是误差矩阵。

可将模型（4.9）看成 $\mathbb{R}^p$ 中的 K 个标准线性回归问题，每个回归问题都享有相同的协变量。$\boldsymbol{\Theta}$ 中的第 k 列 $\theta_k \in \mathbb{R}^p$ 是第 k 个问题的系数向量。因此，理论上可以对 K 个不同问题进行单独拟合来得到回归系数向量 θ_k，在稀疏线性模型中可使用 lasso 方法来拟合。在许多应用中，响应向量 $Y \in \mathbb{R}^K$ 的各个元素是强关联的，所以通常希望回归向量之间最好也有联系。例如，在协同过滤法中，Y 的每个元素表示一位用户对不同类别目标（比如书、电影、音乐等）的喜好程度，它们之间都

有联系。因此，这 K 个回归问题自然可以同时求解，使系数有组结构特征，这样做通常会得到更好的预测结果。再比如，每一个响应值为市场上某只股票的日收益，多支股票以相同的市场信号来预测。

举一个例子，在稀疏情形中，假定有一个未知子集 $S \subset \{1, 2, \ldots, p\}$，其中的协变量与预测相关，且这个子集存在于所有 K 个响应变量中。这个例子自然可以考虑采用一个组 lasso 惩罚，其中通过系数矩阵 $\boldsymbol{\Theta} \in \mathbb{R}^{p \times K}$ 的行 $\{\theta'_j \in \mathbb{R}^K, j = 1, \ldots, p\}$ 来定义 p 个组。为目标函数加上惩罚项，就可求解下面的正则化最小二乘问题：

$$\underset{\boldsymbol{\Theta} \in \mathbb{R}^{p \times K}}{\text{minmize}} \left\{ \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_{\mathrm{F}}^2 + \lambda \left(\sum_{j=1}^{\mathrm{p}} \|\theta'_j\|_2 \right) \right\} \tag{4.10}$$

其中 $\|\cdot\|_{\mathrm{F}}$ 表示 Frobenius 范数。[①] 这个问题是广义组 lasso 式（4.5）的一个特例，其中 $J = p$，对第 j 个组有 $p_j = K$。

4.3.1 组 lasso 计算

下面介绍组 lasso 的计算，我们先用更紧凑的矩阵–向量方式来重写优化问题（4.5）：

$$\underset{(\theta_1, \ldots, \theta_J)}{\text{minmize}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{Z}_j \theta_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \right\} \tag{4.11}$$

为了简单起见，这里忽略了截距 θ_0，因为可对所有的因变量及响应变量做归一化处理，从而去掉截距。对于这个问题，让次梯度为零就会得到如下方程（见 5.2.2 节）：

$$-\boldsymbol{Z}_j^{\mathrm{T}} \left(\boldsymbol{y} - \sum_{\ell=1}^{J} \boldsymbol{Z}_\ell \widehat{\theta}_\ell \right) + \lambda \widehat{s}_j = 0, \quad j = 1, \ldots, J \tag{4.12}$$

其中 $\widehat{s}_j \in \mathbb{R}^{p_j}$ 是在 $\widehat{\theta}_j$ 处范数 $\|\cdot\|_2$ 的次微分。如习题 5.5 所述，如果 $\widehat{\theta}_j \neq 0$，则必然有 $\widehat{s}_j = \widehat{\theta}_j / \|\widehat{\theta}_j\|_2$；如果 $\widehat{\theta}_j = 0$，$\widehat{s}_j$ 则是满足 $\|\widehat{s}_j\|_2 \leqslant 1$ 的任意向量。一种求解零次梯度方程的方法是，固定所有的块向量 $\{\widehat{\theta}_k, k \neq j\}$，然后求解 $\widehat{\theta}_j$。这样做相当于对组 lasso 目标函数采用了块坐标下降法。因为问题是凸的，且惩罚项是块可分的，这就保证可以收敛到一个最优解（Tseng 1993）。固定所有的 $\{\widehat{\theta}_k, k \neq j\}$，则有

$$-\boldsymbol{Z}_j^{T} \left(\boldsymbol{r}_j - \boldsymbol{Z}_j \widehat{\theta}_j \right) + \lambda \widehat{s}_j = 0 \tag{4.13}$$

其中，$\boldsymbol{r}_j = \boldsymbol{y} - \sum_{k \neq j} \boldsymbol{Z}_k \widehat{\theta}_k$ 是第 j 个**偏残差**。由这个与次梯度相关的条件可知：如果

① 对矩阵每个元素简单地取 ℓ_2 范数即可得到该矩阵的 Frobenius 范数。

$\|\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j\| < \lambda$ 则 $\widehat{\theta}_j = 0$，否则最小值 $\widehat{\theta}_j$ 必须满足

$$\widehat{\theta}_j = \left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j + \frac{\lambda}{\|\widehat{\theta}_j\|_2}\boldsymbol{I}\right)^{-1}\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j \tag{4.14}$$

这个迭代公式除了惩罚参数与 $\|\widehat{\theta}_j\|_2$ 有关，其余和岭回归问题的迭代公式类似。遗憾的是，式（4.14）对 $\widehat{\theta}_j$ 没有闭合解，除非 $\boldsymbol{Z}_j$ 标准正交。在这种特殊情况下，迭代公式可以简单化为

$$\widehat{\theta}_j = \left(1 - \frac{\lambda}{\|\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j\|_2}\right)_+ \boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j \tag{4.15}$$

其中，$(t)_+ := \max\{0, t\}$ 是一个函数，若变量为正数，它会返回该量，否则就返回零。习题 4.6 会进一步介绍它。

尽管原作者（Yuan and Lin 2006a）和很多研究者都假设了标准正交，但这并不完全合理（Simon and Tibshirani 2012）。习题 4.8 研究了这样的假设对因子哑变量编码的影响。一般情况下（非正交），式（4.14）需要通过迭代的方法求解，这就可以化简为一维搜索方法（见习题 4.7）。

另一种方法是采用 5.3.3 节的复合梯度方法，这样可以得到一种在每个块中进行迭代的算法。每迭代一次，块优化问题就有一个更简单的近似问题，这样就会得到式（4.15）的迭代公式。迭代公式的具体形式为

$$w \leftarrow \widehat{\theta}_j + \nu \cdot \boldsymbol{Z}_j^{\mathrm{T}}\left(\boldsymbol{r}_j - \boldsymbol{Z}_j\widehat{\theta}_j\right) \tag{4.16a}$$

$$\widehat{\theta}_j \leftarrow \left(1 - \frac{\nu\lambda}{\|w\|_2}\right)_+ w \tag{4.16b}$$

其中 ν 是步长参数。详细推导见习题 4.9。

4.3.2　稀疏组 lasso

当一个组被包含在组 lasso 拟合中时，组中的所有系数都不为零。这是 ℓ_2 范数造成的。有时要依照稀疏性来考虑选择哪些组，还要让组内的系数不为零。例如，虽然某一生物路径和一种特定癌症有关联，但是并非路径上的所有基因都有关联。**稀疏组 lasso**就可以实现组内稀疏性。

为了实现组内稀疏性，可在组 lasso（4.11）上增加一个 ℓ_1 惩罚项，这同样会得到一个凸优化问题

$$\underset{\{\theta_j \in \mathbb{R}^{p_j}\}_{j=1}^J}{\operatorname{minmize}} \left\{\frac{1}{2}\left\|\boldsymbol{y} - \sum_{j=1}^J \boldsymbol{Z}_j\theta_j\right\|_2^2 + \lambda\sum_{j=1}^J\left[(1-\alpha)\|\theta_j\|_2 + \alpha\|\theta_j\|_1\right]\right\} \tag{4.17}$$

其中 $\alpha \in [0,1]$。这与 4.2 节中的弹性网方法非常相似，参数 α 用于在组 lasso（α=0）和 lasso（α=1）之间建立联系。图 4-5 展示了基于 3 个变量的组 lasso 和稀疏组 lasso 的约束区域。注意：在两个水平轴上，约束区域类似于弹性网方法。

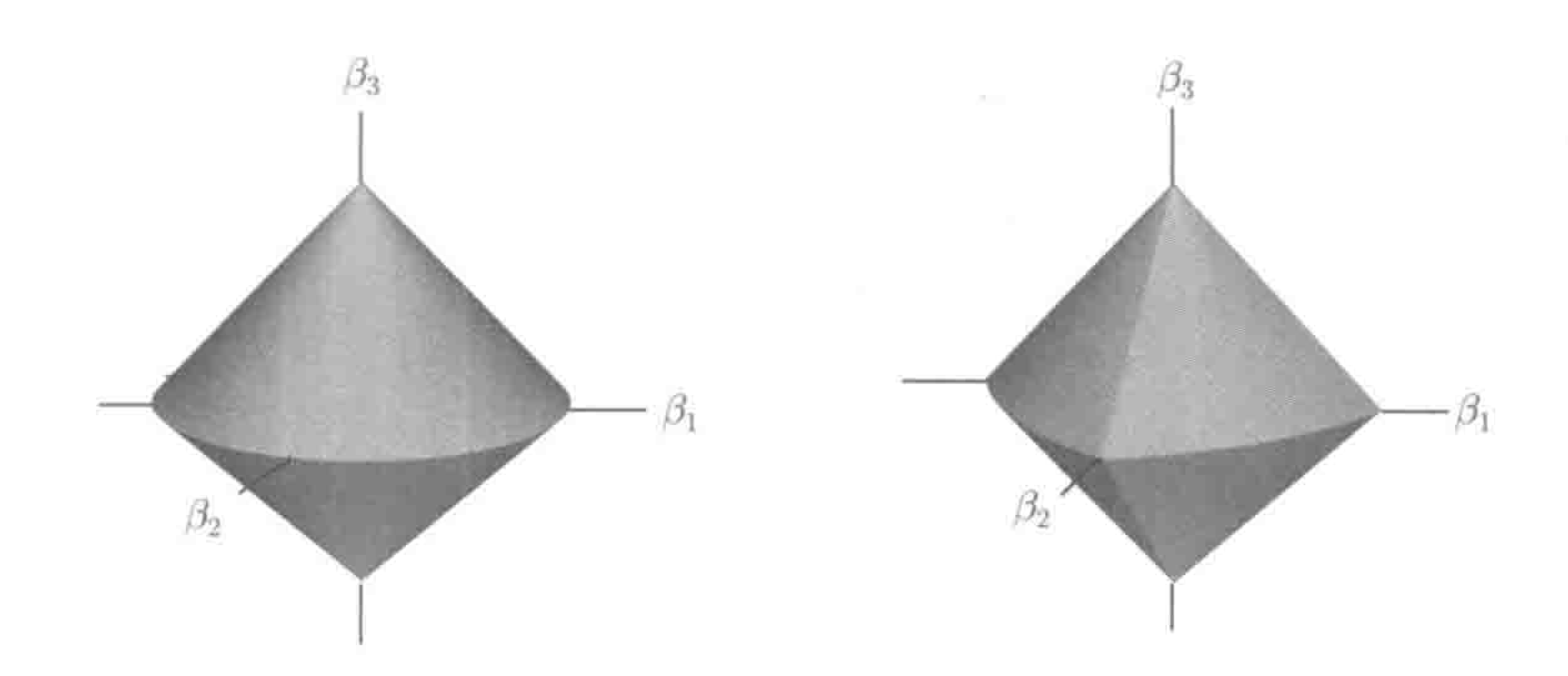


图 4-5 $\mathbb{R}^3$ 中的组 lasso 球（左图）和 α=0.5 下的稀疏组 lasso 球（右图）。这里总共分为两个组，系数分别是 $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$，$\theta_2 = \beta_3 \in \mathbb{R}^1$

因为优化问题（4.17）是凸的，所以最优解满足次梯度方程为零，这类似于组 lasso（4.13）。更准确地讲，任何最优解都必须满足

$$-\boldsymbol{Z}_j^{\mathrm{T}}\left(\boldsymbol{y}-\sum_{\ell=1}^{J}\boldsymbol{Z}_\ell\widehat{\theta}_\ell\right)+\lambda(1-\alpha)\cdot\widehat{s}_j+\lambda\alpha\widehat{t}_j=0, \quad j=1,\ldots,J \tag{4.18}$$

其中 $\widehat{s}_j \in \mathbb{R}^{p_j}$ 是在 $\widehat{\theta}_j$ 处的欧几里得范数的次梯度，$\widehat{t}_j \in \mathbb{R}^{p_j}$ 是在 $\widehat{\theta}_j$处ℓ_1 范数下的次微分。与一般 lasso 一样，$\widehat{t}_{jk} \in \operatorname{sgn}(\theta_{jk})$。

再一次采用块坐标下降法来求解这些方程， 解比之前要复杂一些。 如式（4.13），$\boldsymbol{r}_j$ 为第 j 个坐标下的偏残差，当且仅当方程

$$\boldsymbol{Z}_j^T\boldsymbol{r}_j=\lambda(1-\alpha)\widehat{s}_j+\lambda\alpha\widehat{t}_j \tag{4.19}$$

有解且 $\|\widehat{s}_j\|_2 \leqslant 1$，$\widehat{t}_{jk} \in [-1,1]$，其中 $k=1,\ldots,p_j$。幸运的是，这个条件很容易满足，并可得到如下结论（见习题 4.12）：

$$\widehat{\theta}_j=0, \quad \text{当且仅当}\|S_{\lambda\alpha}(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j)\|_2 \leqslant \lambda(1-\alpha) \tag{4.20}$$

其中 $\mathcal{S}_{\lambda\alpha}(\cdot)$ 是软阈值算子，将向量 $\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j$ 作为该算子的参数，该算子会计算向量的每个分量。注意，这里除了用到软阈值梯度 $\mathcal{S}_{\lambda\alpha}(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j)$，其他与组 lasso（4.13）的情形很相似。同样，如果 $\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j=\boldsymbol{I}$，则有（见习题 4.13）

$$\widehat{\theta}_j=\left(1-\frac{\lambda(1-\alpha)}{\|S_{\lambda\alpha}(\boldsymbol{Z}_j^T\boldsymbol{r}_j)\|_2}\right)_+ S_{\lambda\alpha}(\boldsymbol{Z}_j^T\boldsymbol{r}_j) \tag{4.21}$$

在一般情况下，当 $\boldsymbol{Z}_j$ 之间不正交，且 $\widehat{\theta}_j \neq 0$ 时，求解 $\widehat{\theta}_j$ 等同于求解子问题

$$\underset{\theta_j \in \mathbb{R}^{p_j}}{\operatorname{minmize}} \left\{ \frac{1}{2} \left\| \boldsymbol{r}_j - \boldsymbol{Z}_j \theta_j \right\|_2^2 + \lambda (1-\alpha) \left\| \theta_j \right\|_2 + \lambda\alpha \left\| \theta_j \right\|_1 \right\} \tag{4.22}$$

为了按块来求解，这里会再次采用广义梯度下降法（见 5.3.3 节），得到一个简单的迭代算法，就像式（4.16a）那样。算法会持续迭代直到以下序列收敛：

$$w \leftarrow \widehat{\theta}_j + \nu \cdot \boldsymbol{Z}_j^{\mathrm{T}} \left(\boldsymbol{r}_j - \boldsymbol{Z}_j \widehat{\theta}_j \right) \tag{4.23a}$$

$$\widehat{\theta}_j \leftarrow \left(1 - \frac{\nu\lambda (1-\alpha)}{\left\| S_{\lambda\alpha}(w) \right\|_2} \right)_+ S_{\lambda\alpha}(w) \tag{4.23b}$$

其中 ν 是迭代步长，详见习题 4.10。

4.3.3 重叠组 lasso

有时变量属于多个组，例如，基因可能属于多条生物通路。重叠组 lasso 可以处理变量属于多个组的情况。

为了便于直观理解，下面通过一个例子来说明。考虑将 5 个变量分为两组，即

$$\boldsymbol{Z}_1 = (X_1, X_2, X_3), \quad \boldsymbol{Z}_2 = (X_3, X_4, X_5) \tag{4.24}$$

这里的 X_3 同时属于两个组。重叠组 lasso 模型先将变量复制到它所在的任意组中，然后像之前一样采用普通组 lasso 进行拟和。在这个例子中，X_3 被复制到两个组中，然后采用组 lasso（4.5），其中组惩罚为 $\|\theta_1\|_2 + \|\theta_2\|_2$，由此拟合系数向量 $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ 和$\theta_2 = (\theta_{21}, \theta_{22}, \theta_{23})$。按照原始变量，$X_3$ 的系数 $\widehat{\beta}_3 = \widehat{\theta}_{13} + \widehat{\theta}_{21}$。于是，只要系数 $\widehat{\theta}_{13}$ 和 $\widehat{\theta}_{21}$ 有一个不为零，系数 $\widehat{\beta}_3$ 就不为零。因此，在其他条件相同的情况下，与其他变量相比，变量 X_3 由于同时属于两组，包含在模型中的可能性会更大。

与复制变量相比，在组 lasso 惩罚中简单地复制系数更有效。例如，对于上面给出的组 $X = (X_1, X_2, X_3, X_4, X_5)$ 及 $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$，假设

$$\theta_1 = (\beta_1, \beta_2, \beta_3), \quad \theta_2 = (\beta_3, \beta_4, \beta_5) \tag{4.25}$$

然后像之前一样采用组 lasso，其惩罚项仍为 $\|\theta_1\|_2 + \|\theta_2\|_2$。但是，这个方法有一个明显的缺点：当最优解的 $\widehat{\theta}_1 = 0$ 时，**两个组**都必须使 $\widehat{\beta}_3 = 0$。因此，在本例中，可能的非零系数集合只有$\{1, 2\}$、$\{4, 5\}$和$\{1, 2, 3, 4, 5\}$。原始的分组$\{1, 2, 3\}$和$\{3, 4, 5\}$则不可能存在，因为如果任一一组存在，则两组必须都存在。[①] 第二个重要的问题是：该方法中的惩罚项不可分，因此坐标下降算法可能无法收敛到一个最优解（详见习题 5.4）。

① 更常见情况下，变量复制方法产生的解中，零系数集合通常为组的组合，所以非零系数集合必须是组内元素的交集。

Jacob, Obozinski and Vert（2009）意识到了这个问题，因此提出了复制变量方法（4.24），即重叠组 lasso。对于这里的例子，重叠组 lasso 模型中可能的非零系数集合为{1, 2, 3}、{3, 4, 5}和{1, 2, 3, 4, 5}。可能的非零系数集通常与组（或组的组合）对应。它们也在原始变量上定义隐式的惩罚，这样能得到与复制变量方法一样的解。

符号 $\nu_j \in \mathbb{R}^p$ 表示一个向量，除了与第 j 个组的成员所对应的位置不为零，其余皆为零。设 $\mathcal{V}_j \in \mathbb{R}^p$ 为这些向量的一个子空间。对于原始变量 $X = (X_1, \ldots, X_p)$，系数向量由 $\beta = \sum_{j=1}^{J} \nu_j$ 得到，因此重叠组 lasso 需要求解问题

$$\underset{\nu_j \in \mathcal{V}_j \ j=1,\ldots,J}{\text{minmize}} \left\{ \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{X} \left(\sum_{j=1}^{J} \nu_j \right) \right\|_2^2 + \lambda \sum_{j=1}^{J} \|\nu_j\|_2 \right\} \tag{4.26}$$

定义一个合适的惩罚函数，然后就可以通过最初的 β 变量重写这个优化问题。有了

$$\Omega_{\mathcal{V}}(\beta) := \inf_{\substack{\nu_j \in \mathcal{V}_j \\ \beta = \sum_{j=1}^{J} \nu_j}} \sum_{j=1}^{J} \|\nu_j\|_2 \tag{4.27}$$

就可以证明，求解问题（4.26）等同于求解下面的问题（Jacob et al. 2009）：

$$\underset{\beta \in \mathbb{R}^p}{\text{minmize}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \Omega_{\mathcal{V}}(\beta) \right\} \tag{4.28}$$

这种等价关系很直观，而且强调了这种惩罚的基本原理：这种有效的范数可以将变量系数分散在各个组中。

图 4-6 对比了 3 个变量的情况下，组 lasso 和重叠组 lasso 的约束区域。图中两个环对应两个组，其中变量 X_2 同时属于两组。

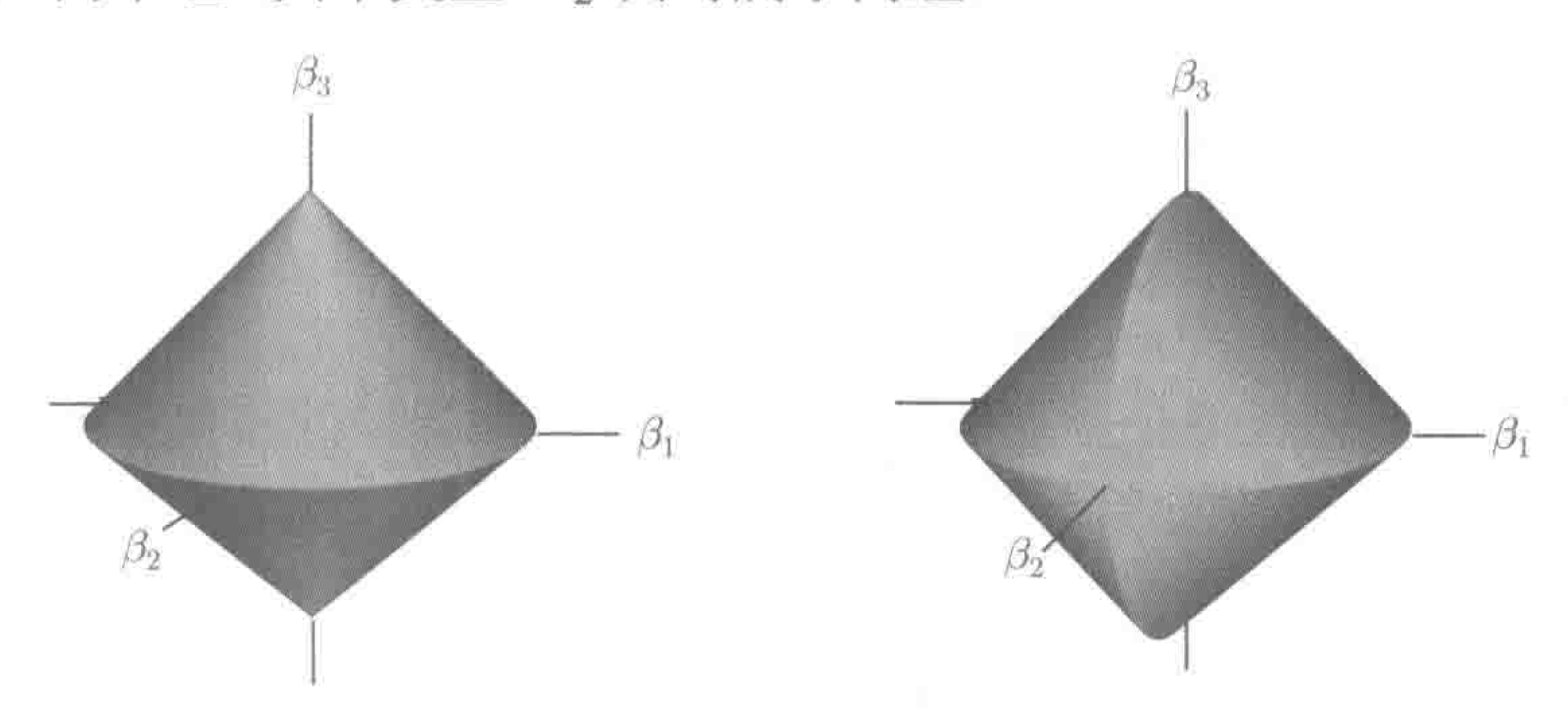


图 4-6　$\mathbb{R}^3$ 中的组 lasso 球（左图）与重叠组 lasso 球（右图）。这里都是分为两组。左图中的组为 $\{X_1, X_2\}$ 和 X_3，右图中的组为 $\{X_1,X_2\}$ 和 $\{X_2,X_3\}$。右图中的两个环对应两组。当 β_2 趋向于零时，对另外两个变量的惩罚与 lasso 一样。当 β_2 远离零时，对其他两个变量的惩罚变“软”了，则类似于 ℓ_2 惩罚

例 4.3：相互作用和分层　要在线性模型中选出相互作用时，重叠组 lasso 模型可以用于**强制分层**。这意味着仅在它们的主效应都存在时，模型中才会有相互作用。假设 Z_1 表示因子 G_1 的 p_1 层中的哑变量 p_1，Z_2 表示 G_2 因子的 p_2 哑变量。有 Z_1 和 Z_2 的线性模型是**主效应**（main-effect）模型。设 $Z_{1:2} = Z_1 * Z_2$，这是一个 $p_1 \times p_2$ 的哑变量向量（成对相乘的向量）。Lim and Hastie（2014）给出了关于分类变量对[①] 的目标函数

$$\begin{aligned}\underset{\mu,\alpha,\widetilde{\alpha}}{\text{min mize}} \left\{ \frac{1}{2} \left\| \boldsymbol{y} - \mu\mathbf{1} - \boldsymbol{Z}_1\alpha_1 - \boldsymbol{Z}_2\alpha_2 - [\boldsymbol{Z}_1 \quad \boldsymbol{Z}_2 \quad \boldsymbol{Z}_{1:2}] \begin{bmatrix} \widetilde{\alpha}_1 \\ \widetilde{\alpha}_2 \\ \alpha_{1:2} \end{bmatrix} \right\|_2^2 \right. \\ \left. +\lambda \left(\|\alpha_1\|_2 + \|\alpha_2\|_2 + \sqrt{p_2 \|\widetilde{\alpha}_1\|_2^2 + p_1 \|\widetilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2} \right) \right\}\end{aligned} \tag{4.29}$$

约束条件为

$$\sum_{i=1}^{p_1} \alpha_1^i = 0, \quad \sum_{j=1}^{p_2} \alpha_2^j = 0, \quad \sum_{i=1}^{p_1} \widetilde{\alpha}_1^i = 0, \quad \sum_{j=1}^{p_2} \widetilde{\alpha}_2^j = 0 \tag{4.30}$$

$$\text{对于固定}j, \quad \sum_{i=1}^{p_1} \alpha_{1:2}^{ij} = 0\text{；对于固定}i, \quad \sum_{j=1}^{p_2} \alpha_{1:2}^{ij} = 0 \tag{4.31}$$

在层次 ANOVA 的公式中，求和的约束很标准。注意，主效应矩阵 Z_1 和 Z_2 各自含有两个不同的系数向量 α_j 和 $\widetilde{\alpha}_j$，在惩罚上有重叠，最终的系数为 $\theta_j = \alpha_j + \widetilde{\alpha}_j$。$\sqrt{p_2\|\widetilde{\alpha}_1\|_2^2 + p_1\|\widetilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2}$ 可得到很强的层次，因为要么 $\widehat{\widetilde{\alpha}}_1 = \widehat{\widetilde{\alpha}}_2 = \widehat{\alpha}_{1:2} = 0$，要么都不为零，例如，相互作用总是与所有主效应同时出现。由此可知，上述约束问题（4.29）～（4.31）的解等价于下面的简单无约束问题的解（见习题 4.14）：

$$\begin{aligned}\underset{\mu,\beta}{\text{minmize}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \mu\mathbf{1} - \boldsymbol{Z}_1\beta_1 - \boldsymbol{Z}_2\beta_2 - \boldsymbol{Z}_{1:2}\beta_{1:2}\|_2^2 \right. \\ \left. +\lambda (\|\beta_1\|_2 + \|\beta_2\|_2 + \|\beta_{1:2}\|_2) \right\}\end{aligned} \tag{4.32}$$

换句话说，带 $Z_{1:2}$ 的模型是全相互作用模型（即，与主效应的相互作用已经包含在内）。含 Z_1、Z_2 和 $Z_{1:2}$ 的组 lasso 模型会得出分层模型。当 $Z_{1:2}$ 在模型中时，主效应对总是隐式地包含在内。在这种情况下，变量不是严格重叠，但它们的子空间却是。Bien, Taylor and Tibshirani（2013）提出了 hierNet 方法，这是一种估计分层交互作用的方法。

---

① 这可以自然地扩展到两对以上，同样可以扩展到其他的损失函数（例如，逻辑斯蒂回归），也可以扩展到因子和定量变量间的相互作用。

4.4 稀疏加法模型和组 lasso

假设有一个零均值响应变量 $Y \in \mathbb{R}$，还有一个预测子向量 $X \in \mathbb{R}^J$，要估计回归方程 $f(x) = \mathbb{E}(Y \mid X = x)$。众所周知，非参数回归因维数灾难而饱受诟病，因此近似法至关重要。加法模型正是这样的近似方法，它能够有效地将估计问题转化为一维问题。如果 J 非常大，这可能还不够。稀疏加法模型将某些成分变成零，进一步限制这些近似，以促使更多系数为零来更多地限制这种近似。稀疏加法模型的估计方法与组 lasso 相关。

4.4.1 加法模型和 backfitting

这里介绍一下加法模型的背景。加法模型以求和的形式来近似回归函数：

$$f(x) = f(x_1, \ldots, x_J) \approx \sum_{j=1}^{J} f_j(x_j)$$
$$f_j \in \mathcal{F}_j,\ j = 1, \ldots, J \tag{4.33}$$

其中 $\mathcal{F}_j$ 是给定集合上的单变量函数。通常假设每个 $\mathcal{F}_j$ 都是 $L^2(\mathbb{P}_j)$ 的一个子集，其中 $\mathbb{P}_j$ 是协变量 X_j 的分布，其 $L^2(\mathbb{P}_j)$ 范数为 $\|f_j\|_2^2 := \mathbb{E}[f_j^2(X_j)]$。

在一般情形下，采用 $L^2(\mathbb{P})$ 作为度量，近似回归函数 $\mathbb{E}(Y \mid X = x)$ 的最佳加法模型可以求解下面的问题：

$$\underset{f_j \in \mathcal{F}_j, j=1,\ldots,J}{\text{minimize}} \mathbb{E}[(Y - \sum_{j=1}^{J} f_j(X_j))^2] \tag{4.34}$$

最优解 $(\widetilde{f}_1, \ldots, \widetilde{f}_J)$ 可由 backfitting 方程来刻画，即

$$\widetilde{f}_j(x_j) = \mathbb{E}[Y - \sum_{k \neq j} \widetilde{f}_k(X_k) \mid X_j = x_j], \quad j = 1, \ldots, J \tag{4.35}$$

迭代公式有更紧凑的形式 $\widetilde{f}_j = \mathcal{P}_j(R_j)$，其中 $\mathcal{P}_j$ 是在第 j 个坐标中的条件期望算子，变量 $R_j := Y - \sum_{k \neq j} \widetilde{f}(X_k)$ 是第 j 个偏残差。

给定数据 $\{(x_i, y_i)\}_{i=1}^{N}$，一个自然的方法就是用经验版本（例如散点图滤波算子 $\mathcal{S}_j$）来替代总体算子 $\mathcal{P}_j$，然后通过坐标下降法或者 backfitting 法来求解基于数据版本的更新式（4.35）（Hastie and Tibshirani 1990）。因此可按坐标 $j = 1, \ldots, J$ 进行循环，然后用平滑的偏残差来更新各个估计公式 $\widehat{f}_j$：

$$\widehat{f}_j \leftarrow \mathcal{S}_j \left(\boldsymbol{y} - \sum_{k \neq j} \widehat{\boldsymbol{f}}_k \right), \quad j = 1, \ldots, J \tag{4.36}$$

直到拟合后的方程 $\widehat{f}_j$ 稳定。在式（4.36）中，$\widehat{\boldsymbol{f}}_k$ 是拟合后的方程 $\widehat{f}_k$ 在 N 个样本 $(x_{1k}, \ldots, x_{Nk})$ 上的估计值。算子 $\mathcal{S}_j$ 代表算法，该算法的输入为响应向量 $\boldsymbol{r}$，根据

向量 $\boldsymbol{x}_j$ 对其进行平滑处理，并返回方程 $\widehat{f}_j$。尽管算子 $\mathcal{S}_j$ 有自己的可调参数，但此时可将其当成一个黑盒，使用数据来估计条件期望。

4.4.2　稀疏加法模型和 backfitting

稀疏加法模型是基本加法模型的扩展，可假设有一个子集 $S \subset \{1,2,\ldots,J\}$，回归函数 $f(x)=\mathbb{E}(Y\,|X=x)$ 满足 $f(x)\approx\sum_{j\in S}f_j(x_j)$ 形式的一个近似。Ravikumar, Liu, Lafferty and Wasserman（2009）受总体水平问题（4.34）的启发，提出 backfitting 方程的一个扩展。对于一个给定的稀疏层级 $k\in\{1,\ldots,J\}$，回归函数的最优 k 稀疏近似为

$$\underset{\substack{|S|=k\\ f_j\in\mathcal{F}_j,j=1,\ldots,J}}{\text{minimize}}\ \mathbb{E}\left(Y-\sum_{j\in S}f_j(X_j)\right)^2 \tag{4.37}$$

遗憾的是，这是一个非凸问题，而且会涉及一个与 k 的大小相关的组合数 $\binom{J}{k}$，因此问题难以计算。假设换一种方法，通过 $\sum_{j=1}^J\|f_j\|_2$ 得到加法近似 $f=\sum_{j=1}^J f_j$ 的稀疏性，其中 $\|f_j\|_2=\sqrt{\mathbb{E}[f_j^2(X_j)]}$ 是将 $L^2(\mathbb{P})$ 范数用到第 j 个分量上。对于给定的正则参数 $\lambda\geqslant 0$，这种松弛方法可以替代最优稀疏近似，即最小化下面的目标函数：

$$\underset{f_j\in\mathcal{F}_j,j=1,\ldots,J}{\text{minimize}}\left\{\mathbb{E}\left(Y-\sum_{j=1}^J f_j(X_j)\right)^2+\lambda\sum_{j=1}^J\|f_j\|_2\right\} \tag{4.38}$$

该目标函数 $(f_1,\ldots,f_J)$ 是凸函数，拉格朗日对偶确保它有一个等价的表示，涉及范数 $\sum_{j=1}^J\|f_j\|_2$ 上的显示约束（见习题 4.15）。

Ravikumar et al.（2009）证明了式（4.38）的任意最优解 $(\widetilde{f}_1,\ldots,\widetilde{f}_J)$ 都可由**稀疏 backfitting 方程**刻画：

$$\widetilde{f}_j=\left(1-\frac{\lambda}{\|\mathcal{P}_j\left(R_j\right)\|_2}\right)_+\mathcal{P}_j\left(R_j\right) \tag{4.39}$$

其中，残差 R_j 和条件期望算子 $\mathcal{P}_j$ 的定义在式（4.35）之后。

与早先的研究一样，给定数据 $\{(x_i,y_i)\}_1^N$，从这些总体水平更新公式可看出这是对数据驱动的模拟，其中散点图平滑器 $\mathcal{S}_j$ 替换了总体算子 $\mathcal{P}_j$，然后按下面的公式来进行更新：

$$\widetilde{f}_j=S_j\left(\boldsymbol{y}-\sum_{k\neq j}\widehat{\boldsymbol{f}}_k\right),\quad \widehat{f}_j=\left(1-\frac{\lambda}{\|\widetilde{\boldsymbol{f}}\|_2}\right)_+\widetilde{f}_j \tag{4.40}$$

沿 $j=1,\ldots,J$ 进行迭代，直至收敛。图 4-7 展示了 SPAM 迭代（4.40）在一些空气污染数据上的表现。这里采用了光滑样条，各维坐标的自由度固定为 df=5（Hastie and Tibshirani 1990）。

如果将对变量 X_j 的光滑方法投影在一组基函数上，则可以与组 lasso 建立直接的联系。考虑

$$f_j(\cdot)=\sum_{\ell=1}^{p_j}\psi_{j\ell}(\cdot)\beta_{j\ell} \tag{4.41}$$

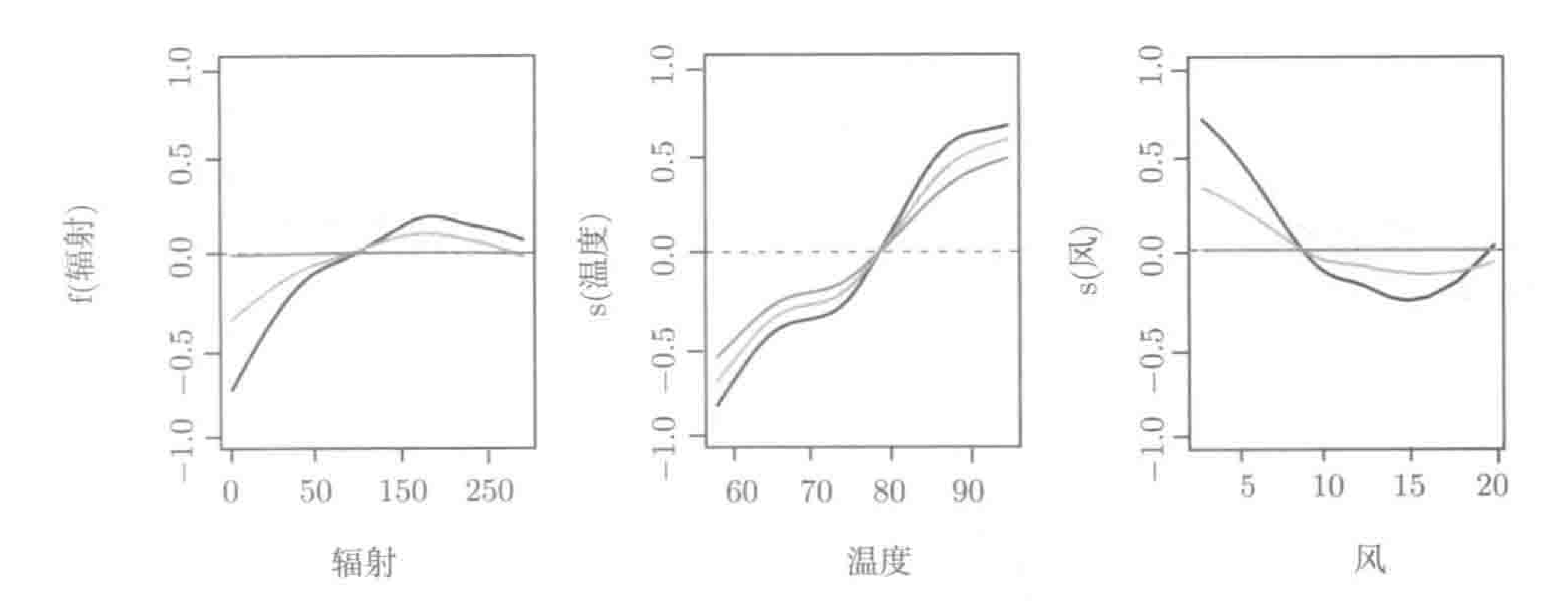


图 4-7　对空气污染数据，采用三个 SPAM 模型进行拟合所得的结果。响应变量是臭氧浓度取对数，有三个预测子：辐射、温度和风速。在拟合加法模型中用到了光滑样条，其 $df=5$。图中三条曲线分别对应 λ=0（黑线），λ=2（橙线），和 λ=4（红线）。可以看到，收缩使得温度函数相对不受影响，辐射和风速方面则影响显著（见彩插）

其中 $\{\psi_{j\ell}\}_1^{p_j}$ 是 X_j 上的一组基函数，例如 X_j 范围内节点集合的三次样条。设 $\boldsymbol{\psi}_j$ 为一个 $N\times p_j$ 矩阵，它是对 $\psi_{j\ell}$ 的评估，假设 $\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\psi}_j=I_{pj}$。那么，对于任意系数向量 $\theta_j=\left(\beta_{j1},\ldots,\beta_{jp_j}\right)^{\mathrm{T}}$ 和相应的拟合向量 $\boldsymbol{f}_j=\boldsymbol{\psi}_j\theta_j$，有 $\|\boldsymbol{f}_j\|_2=\|\theta_j\|_2$。在这种情况下，很容易证明更新式（4.40）等价于组 lasso 模型迭代公式（详见习题 4.16）。该模型基于预测子矩阵 $\boldsymbol{\psi}:=[\boldsymbol{\psi}_1\quad\boldsymbol{\psi}_2\quad\ldots\quad\boldsymbol{\psi}_J]$ 和相应系数块向量 $\theta:=[\theta_1\quad\theta_2\quad\ldots\quad\theta_J]$。

4.4.3 优化方法与组 lasso

虽然总体水平稀疏 backfitting 方程（4.39）可以求解优化问题，但在一般情况下，经版式（4.40）不这样做，而是通过类比总体版来得到启发。下面来讨论成分选择和光滑算子（Component Selection and Smoothing Operator，COSSO），这可以求解数据定义（data-define）的优化问题。COSSO 方法（Lin and Zhang 2003）是 SPAM 方法的前身，它基于再生核希尔伯特空间，是光滑样条模型的特例。

先回忆一下加法光滑-样条模型的形式，这个形式可以通过优化下面的目标函数得到：

$$\underset{f_j \in \mathcal{H}_j, j=1,\ldots,J}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{J} f_j(x_{ij}) \right)^2 + \lambda \sum_{j=1}^{J} \frac{1}{\gamma_j} \|f_j\|_{\mathcal{H}_j}^2 \right\} \tag{4.42}$$

其中，$\|f_j\|_{\mathcal{H}_j}$ 是第 j 个坐标上某个基于希尔伯特空间的范数。通常选择希尔伯特空间 $\mathcal{H}_j$ 是为了实现某种形式的光滑。这种情况下，参数 $\lambda \geqslant 0$ 对应全局光滑，参数 $\gamma_j \geqslant 0$ 是具体坐标的调整器。例如，[0,1] 区间的三次光滑样条的范数为

$$\|g\|_{\mathcal{H}}^2 := \left(\int_0^1 g(t)\,\mathrm{d}t \right)^2 + \left(\int_0^1 g'(t)\,\mathrm{d}t \right)^2 + \int_0^1 g''(t)^2\,\mathrm{d}t \tag{4.43}$$

在目标函数式（4.42）上采用这种希尔伯特范数时，最优解中每个分量 $\widehat{f}_j$ 都是在 X_j 处的带节点三次样条。通过 backfitting 迭代（4.36）可得到解，其中 $\mathcal{S}_j$ 是一类具有惩罚 λ/γ_j 的三次样条光滑函数

与经典式（4.42）不同，COSSO 方法的目标函数为

$$\underset{f_j \in \mathcal{H}_j, j=1,\ldots,J}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{J} f_j(x_{ij}) \right)^2 + \tau \sum_{j=1}^{J} \|f_j\|_{\mathcal{H}_j} \right\} \tag{4.44}$$

如前一样，惩罚项是范数而非范数的平方，足够大的 τ 会引发坐标选择。注意，与三次光滑样条的惩罚项不同，式（4.43）中的范数包含了线性成分。这确保了这一项排除在模型之外后，整个函数为零，而不是只有非线性成分。虽然这与加法样条问题（4.38）相似，但惩罚项 $\|f_j\|_{\mathcal{H}_j}$ 的结构意味着，解并不像稀疏 backfitting 方程（4.40）那样简单。

采用范数（4.44），三次样条空间 $\mathcal{H}_j$ 是再生核希尔伯特空间（Reproduling Kernel Hilbert Space, RKHS）在单位区间 [0,1] 上的一个特例。这种空间可由对称正定核函数 $\mathcal{R}_j : [0,1] \times [0,1] \to \mathbb{R}$ 来刻画，这类函数具有所谓的再生性。具体而言，对任意 $x \in [0,1]$，可保证函数 $\mathcal{R}_j(\cdot, x)$ 属于 $\mathcal{H}_j$，而且对所有的 $f \in \mathcal{H}_j$，$\langle \mathcal{R}_j(\cdot, x), f \rangle_{\mathcal{H}_j} = f(x)$。这里的 $\langle \cdot, \cdot \rangle_{\mathcal{H}_j}$ 表示在希尔伯特空间 $\mathcal{H}_j$ 上的内积。

运用再生性，可证明（见习题 4.17）：对于 COSSO 的任意最优解，若选择合适的权重向量 $\widehat{\boldsymbol{\theta}}_j \in \mathbb{R}^N$，第 j 个坐标函数 $\widehat{f}_j$ 可以写成 $\widehat{f}_j(\cdot) = \sum_{i=1}^{N} \widehat{\theta}_{ij} \mathcal{R}_j(\cdot, x_{ij})$。此外，可以证明 $\widehat{f}_j$ 在希尔伯特空间的范数 $\|\widehat{f}_j\|_{\mathcal{H}_j}^2 = \widehat{\boldsymbol{\theta}}_j^{\mathrm{T}} \boldsymbol{R}_j \widehat{\boldsymbol{\theta}}_j$，其中 $\boldsymbol{R}_j \in \mathbb{R}^{N \times N}$ 是一个由核定义的**Gram 矩阵**。具体而言，每一项 $(\boldsymbol{R}_j)_{ii'} = \mathcal{R}_j(x_{ij}, x_{i'j})$。因此，COSSO 问题（4.44）可重写为一个更为常见的组 lasso 形式，实际上它等价于优化问题

$$\underset{\boldsymbol{\theta}_j \in \mathbb{R}^N, j=1,\ldots,J}{\text{minimize}} \left\{ \frac{1}{N} \left\| \boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{R}_j \boldsymbol{\theta}_j \right\|_2^2 + \tau \sum_{j=1}^{J} \sqrt{\boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{R}_j \boldsymbol{\theta}_j} \right\} \tag{4.45}$$

证明见习题 4.17。

现在来讨论参数设定，其解是组 lasso（4.14）更一般的版本。可以证明，任意的最优解 $(\widehat{\boldsymbol{\theta}}_1,\ldots,\widehat{\boldsymbol{\theta}}_J)$ 可由不动点（fixed point）方程

$$\widehat{\boldsymbol{\theta}}_j=\begin{cases}\mathbf{0}, & \sqrt{\widehat{\boldsymbol{\theta}}_j^{\mathrm{T}}\boldsymbol{R}_j\widehat{\boldsymbol{\theta}}_j}<\tau\\ \left(\boldsymbol{R}_j+\dfrac{\tau}{\sqrt{\widehat{\boldsymbol{\theta}}_j^{\mathrm{T}}\boldsymbol{R}_j\widehat{\boldsymbol{\theta}}_j}}\boldsymbol{I}\right)^{-1}\boldsymbol{r}_j, & \text{其他}\end{cases} \tag{4.46}$$

得到，其中 $\boldsymbol{r}_j=\boldsymbol{y}-\sum_{k\neq j}\boldsymbol{R}_k\widehat{\boldsymbol{\theta}}_k$ 为第 j 个偏残差。尽管 $\widehat{\boldsymbol{\theta}}_j$ 同时出现在式（4.46）的两端，但是对 $\boldsymbol{R}_j$ 进行一次 SVD 分解和简单的一维搜索就能求解。详见习题 4.7。

Lin and Zhang（2003）提出了一种替代方法，需要引入一个辅助向量 $\boldsymbol{\gamma}\in\mathbb{R}^J$，然后求解联合优化函数

$$\underset{\substack{\gamma\geqslant 0\\ f_j\in\mathcal{H}_j,j=1,\ldots,J}}{\operatorname{minimize}}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{J}f_j\left(x_{ij}\right)\right)^2+\sum_{j=1}^{J}\frac{1}{\gamma_j}\|f_j\|_{\mathcal{H}_j}^2+\lambda\sum_{j=1}^{J}\gamma_j\right\} \tag{4.47}$$

如习题 4.18 所示，如果在式（4.47）中设定 $\lambda=\tau^4/4$，则其最优解的 $\widehat{\boldsymbol{\theta}}$ 也是原 COSSO 问题（4.44）的最优解。

重写的式（4.47）十分有用，由此可以很自然地得到一种有效的算法，交替执行以下两步。

- 固定 γ_j，问题就变成了目标函数（4.42）的形式，这是一个加法样条函数拟合。
- 固定拟合好的加法样条，更新系数向量 $\gamma=(\gamma_1,\ldots,\gamma_J)$，可以得到一个非负的 lasso 问题。更确切地说，对每一个 $j=1,\ldots,J$，需定义向量 $\boldsymbol{g}_j=\boldsymbol{R}_j\boldsymbol{\theta}_j/\gamma_j\in\mathbb{R}^N$，其中 $\boldsymbol{f}_j=\boldsymbol{R}_j\boldsymbol{\theta}_j$ 是采用当前值 γ_j 对第 j 个函数拟合而得的向量。然后通过求解目标函数

$$\min_{\gamma\geqslant 0}\left\{\frac{1}{N}\|\boldsymbol{y}-\boldsymbol{G}\gamma\|_2^2+\lambda\|\gamma\|_1\right\} \tag{4.48}$$

  来更新向量 $\gamma=(\gamma_1,\ldots,\gamma_J)$，其中 $\boldsymbol{G}$ 是一个 $N\times J$ 矩阵，每一列为 $\{\boldsymbol{g}_j,j=1,\ldots,J\}$。这些更新公式与 Lin and Zhang（2003）给出的有一点不同，详见习题 4.19。

采用三次光滑样条范数式（4.43）时，COSSO 问题令函数 f_j 为零。这个基本思想可以多加拓展。例如，给定一个单变量函数 g，可以用 $g(t)=\alpha_0+\alpha_1 t+h(t)$ 来代替各个单变量函数，采用以下范数可使得惩罚项不是线性的：

$$\|h\|_{\mathcal{H}}^2:=\int_0^1 h''(t)^2\,\mathrm{d}t \tag{4.49}$$

在这种情形下，COSSO 可以对每个分量函数是否为线性作出选择。

4.4.4 节会更加深入地讨论加法模型的惩罚项，实际上，在这种情况下采用多种惩罚项是有益处的。

4.4.4 稀疏加法模型的多重惩罚

目前为止，非参数模型已经有多种方法可以实现稀疏性了。一些方法（如 SPAM backfitting 方法）基于 ℓ_1 范数和经验 ℓ_2 范数的组合，即

$$\|f\|_{N,1} := \sum_{j=1}^{J} \|f_j\|_N \tag{4.50}$$

其中 $\|f_j\|_N^2 := \frac{1}{N}\sum_{i=1}^{N} f_j^2(x_{ij})$ 是单变量函数 f_j 的平方经验 ℓ_2 范数。[①] 其他的方法，例如 COSSO 方法，通过将 ℓ_1 范数与希尔伯特空间的范数相结合

$$\|f\|_{\mathcal{H},1} := \sum_{j=1}^{J} \|f_j\|_{\mathcal{H}_j} \tag{4.51}$$

来实现稀疏性。在非参数情形中，这两种正则化方法中的哪一种能得到更好的稀疏性呢？

除了考虑一种正则项，也可以考虑更一般的估计族

$$\min_{\substack{f_j\in\mathcal{H}_j\\ j=1,\dots,J}} \left\{ \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{J} f_j(x_{ij})\right)^2 + \lambda_{\mathcal{H}}\sum_{j=1}^{J}\|f_j\|_{\mathcal{H}_j} + \lambda_N \sum_{j=1}^{J}\|f_j\|_N \right\} \tag{4.52}$$

式中参数为一对非负正则化权重 $(\lambda_{\mathcal{H}}, \lambda_N)$。如果设定 $\lambda_N = 0$，则优化问题（4.52）会变成 COSSO 估计；如果 $\lambda_{\mathcal{H}} = 0$，则得到一个类似于 SPAM 估计的方法。对于任意非负 $(\lambda_{\mathcal{H}}, \lambda_N)$，优化问题（4.52）是凸的。当 $\mathcal{H}_j$ 为再生核希尔伯特空间时，问题（4.52）可以重写成一个二阶锥规划，这与组 lasso 十分相近。当希尔伯特空间 $\mathcal{H}_j$ 由一个再生核 $\mathcal{R}_j$ 来定义时，对于权重向量 $\widehat{\theta}_j \in \mathbb{R}^N$，任意最优解的第 j 个坐标函数 $\widehat{f}_j$ 都为 $\widehat{f}_j(\cdot) = \sum_{i=1}^{N} \widehat{\theta}_{ij}\mathcal{R}_j(\cdot, x_{ij})$。这就将无限维问题（4.52）转换成为简单问题

$$\min_{\substack{\boldsymbol{\theta}_j\in\mathbb{R}^N\\ j=1,\dots,j}} \left\{ \frac{1}{N}\left\|y - \sum_{j=1}^{J}\boldsymbol{R}_j\boldsymbol{\theta}_j\right\|_2^2 + \lambda_{\mathcal{H}}\sum_{j=1}^{J}\sqrt{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{R}_j\boldsymbol{\theta}_j} + \lambda_N\sum_{j=1}^{J}\sqrt{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{R}_j^2\boldsymbol{\theta}_j} \right\} \tag{4.53}$$

就像前面一样，对任意坐标 $j \in \{1, \dots, J\}$，矩阵 $\boldsymbol{R}_j \in \mathbb{R}^{N\times N}$ 是核 Gram 矩阵，矩阵值为 $[\boldsymbol{R}_j]_{ii'} = \mathcal{R}_j(x_{ij}, x_{i'j})$。详见习题 4.20。

① $\|f_j\|_{N,1}$ 与 4.4.2 节中的 $\|\boldsymbol{f}_j\|_2$ 相同，这里用了更一般的符号。

优化问题（4.53）是一个二阶锥规划例子，可以通过之前描述过的方法的变体来进行求解。但是为什么采用两个正则项就有帮助呢？原因在于，这两个正则项结合在一起会得到极小极大–最优估计，在这种估计下有可能出现最快的收敛率，该收敛率是样本大小、问题维度和稀疏性的函数 [Raskutti, Wainwright and Yu（2012）]。

4.5 融合 lasso

图 4-8 中的灰色尖刺，是比较基因组（comparative genomichybridization，CGH）杂交实验的结果。各个尖刺代表了（log 以 2 为底）肿瘤样本相对于对照组样本的基因副本数，根据基因的染色体顺序画出。这些数据中有很多噪声，所以一定要进行光滑处理。从生物学考虑，要复制的是典型的染色体分段而不是单个基因。因此，真实副本数的向量在染色体的连续区域上可能需要分段恒定。**融合 lasso 信号估计**会提取信号的这种结构，这需要求解优化问题

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\theta_i)^2+\lambda_1\sum_{i=1}^{N}|\theta_i|+\lambda_2\sum_{i=2}^{N}|\theta_i-\theta_{i-1}|\right\} \tag{4.54}$$

第一个惩罚项是常见的 ℓ_1 范数，其作用在于让 θ_i 趋向于零。样本索引 i 针对的是排好序的数据（本例会沿用染色体排列），第二个惩罚项促使相邻系数 θ_i 趋于相同，而且会导致一些系数一样 [即全变分（total-variation）去噪]。注意，式（4.54）并不包含常数项 θ_0，系数 θ_i 直接表示响应变量 y_i，对于这一类问题，零是一个自然起点，详见习题 4.21。图 4-8 中的粗线是用融合 lasso 来拟合这些数据而得到的。

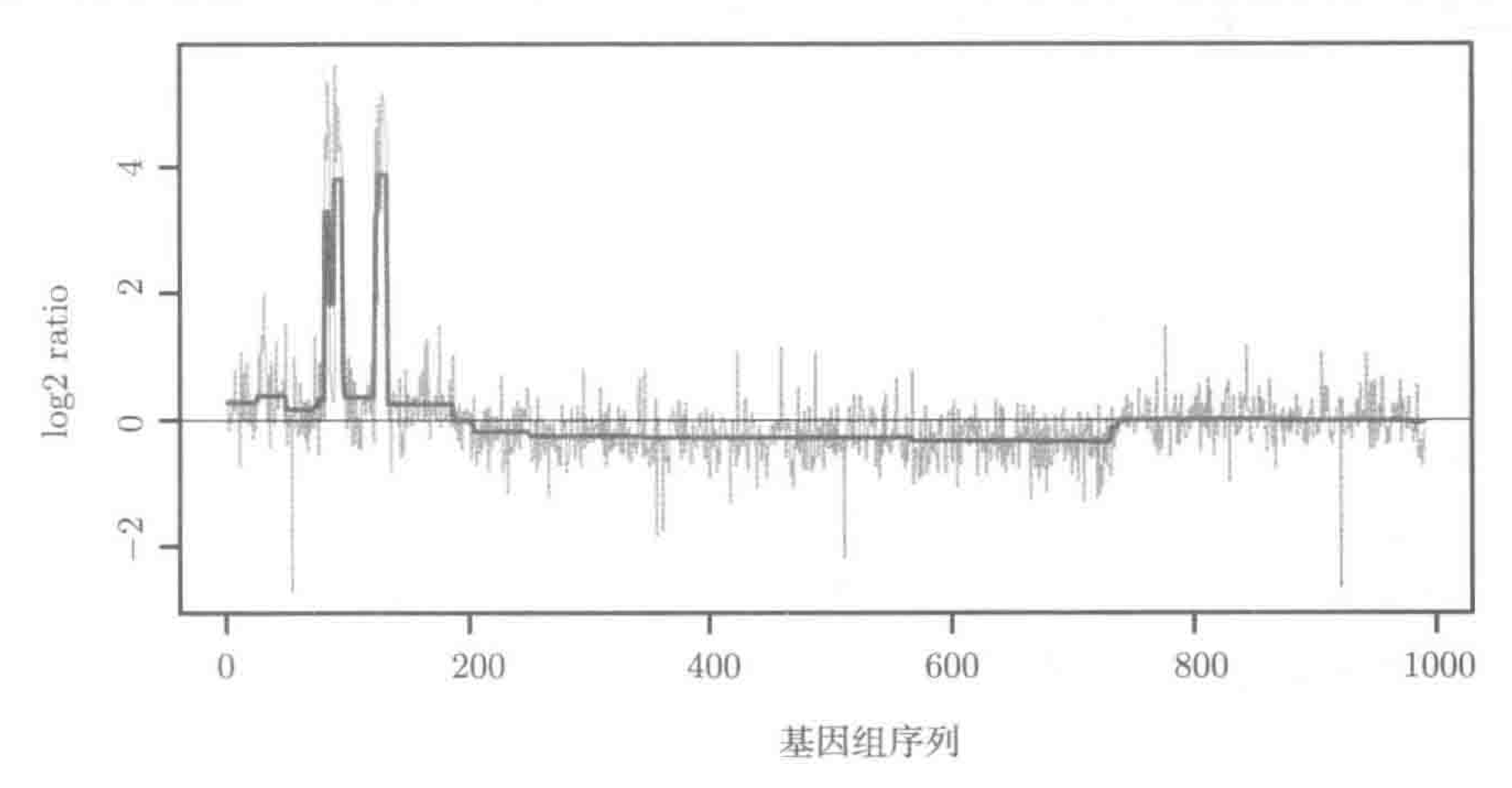


图 4-8 融合 lasso 用于 CGH 数据。每个尖刺代表一个肿瘤样本的副本数，同对照组（在 log 以 2 为底的尺度下）相对应。图中分段的粗线是融合 lasso 估计值

融合 lasso 有很多更一般的形式，这里介绍两个。

- 将相邻的概念从线性排序推广到更一般的相邻关系，如图像中的相邻像素。这就引出了惩罚方式

$$\lambda_2 \sum_{i \sim i'} |\theta_i - \theta_{i'}| \tag{4.55}$$

  这里会对所有的相邻对 $i \sim i'$ 求和。

- 在式（4.54）中，每一个样本值都和一个系数相关联。更一般而言，需要求解目标函数

$$\underset{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}|\beta_j - \beta_{j-1}| \right\} \tag{4.56}$$

  这里协变量 x_{ij} 及系数 β_j 按照某种序列 j 进行索引，使相邻聚集有意义。式（4.54）显然是一个特例。

4.5.1 拟合融合 lasso

式（4.54）及其变体均是凸优化问题，因此它们都能很好地求解。就像其他问题一样，为了在可调参数的取值范围内求解，要寻找一种有效的**路径算法**（path algorithm）。虽然坐标下降法常用于求解 lasso 类问题，但它不能用来求解融合 lasso 问题（4.54），因为各个惩罚项对每个坐标而言并不是一个可分离函数。因此，坐标下降法会像图 5-8 中那样 “停” 在一个非最优点上。这种可分性条件将在 5.4 节中详细讨论。

这里将融合 lasso 问题（4.54）的最优解 $\widehat{\boldsymbol{\theta}}(\lambda_1, \lambda_2)$ 的结构视为两个正则化参数 λ_1 和 λ_2 的函数。Friedman et al.（2007）深入研究了这种最优化的行为，并提出了如下结果。

引理 4.1　对于任意的 $\lambda_1' > \lambda_1$，有

$$\widehat{\theta}_i(\lambda_1', \lambda_2) = \mathcal{S}_{\lambda_1' - \lambda_1}\left(\widehat{\theta}_i(\lambda_1, \lambda_2)\right), \quad i = 1, \ldots, N \tag{4.57}$$

其中 $\mathcal{S}$ 是软阈值算子 $\mathcal{S}_\lambda(z) := \text{sgn}(z)(|z| - \lambda)_+$。

引理 4.1 的一个重要的特例就是等式

$$\widehat{\theta}_i(\lambda_1, \lambda_2) = \mathcal{S}_{\lambda_1}\left(\widehat{\theta}_i(0, \lambda_2)\right), \quad i = 1, \ldots, N \tag{4.58}$$

因此，如果设 $\lambda_1 = 0$，然后求解融合 lasso 问题，那么所有其他的解都可以通过软阈值立即得到。这种有用的变换也可应用到融合 lasso 问题（4.55）的其他更一般的版本中。基于引理 4.1，可以求解问题 ①

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\text{minimize}} \left\{ \frac{1}{2}\sum_{i=1}^{N}(y_i - \theta_i)^2 + \lambda \sum_{i=2}^{N}|\theta_i - \theta_{i-1}| \right\} \tag{4.59}$$

① 这里用符号 λ（而不是 λ_2）来表示正则化参数，因为这里只有一个惩罚项。

下面介绍求解式（4.59）的方法。

1. 重新参量化

一个简单的方法就是重新参数化问题（4.59），让惩罚项具有可加性。具体而言，假设对可逆矩阵 $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ 进行 $\boldsymbol{\gamma} = \boldsymbol{M}\boldsymbol{\theta}$ 形式的线性变换，即

$$\gamma_1 = \theta_1, \quad \gamma_i = \theta_i - \theta_{i-1}, \quad i = 2, \ldots, N \tag{4.60}$$

在这些变换后的坐标中，问题（4.59）等价于普通的 lasso 问题

$$\underset{\gamma\in\mathbb{R}^N}{\operatorname{minimize}}\left\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\gamma}\|^2+\lambda\|\gamma\|_1\right\}, \quad 其中\boldsymbol{X}=\boldsymbol{M}^{-1} \tag{4.61}$$

重新参数化问题（4.61）可以采用任何对 lasso 有效的算法来求解，包括坐标下降法、投影梯度下降法或 LARS 算法。但是，$\boldsymbol{X}$ 是一个下三角矩阵，所有的非零元素均为 1，所以该“变量”具有很大的相关性。无论是坐标下降法还是 LARS，在这种情形下表现都不好（见习题 4.22）。所以尽管重新参数化看起来解决了问题，但是并不值得推荐，其实还有更有效的算法。

2. 路径算法

一维融合 lasso 问题（4.59）有一个有趣的性质：随着正则化参数 λ 的增加，最优解片段只能连接到一起，不能分开。更准确地说，如果设 $\widehat{\boldsymbol{\theta}}(\lambda)$ 表示凸问题（4.59）的最优解，它是 λ 的函数，则有：

引理 4.2: 单调融合 假设对一些 λ 值和一些序列 $i \in \{1, \ldots, N-1\}$，最优解满足 $\widehat{\theta}_i(\lambda) = \widehat{\theta}_{i+1}(\lambda)$，则对所有的 $\lambda' > \lambda$，还有 $\widehat{\theta}_i(\lambda') = \widehat{\theta}_{i+1}(\lambda')$。

该引理的发现极大地简化了融合 lasso 解的路径的构建（Friedman et al. 2007）。可以从 λ=0 开始，这时没有融合组，然后计算能形成融合组的 λ 的最小值。对路径的剩余部分，这个组的估计参数会融合在一起（例如，约束至相等）。采用这种方法，对各个融合组内的估计可以采用一个简单的公式，所以这种方法非常快，只需要 $\mathcal{O}(N)$ 次操作。但要注意：引理 4.2 中的单调融合特性是一维融合 lasso 所特有的。例如，这对具有模型矩阵 $\boldsymbol{X}$ 的广义融合 lasso（4.56）并不成立，对二维融合 lasso（4.55）也不成立。详见 Friedman et al.（2007）以及 Hoefling（2010）。

3. 对偶路径算法

还有一种方法可以得到融合 lasso 问题的**凸对偶**的路径算法（Tibshirani$_2$ and Taylor 2011）。这里通过式（4.59）来解释这个方法，需注意，这个方法也可以运用在广义问题（4.56）上。

问题（4.59）可通过等价的提升形式重写成目标函数

$$\underset{(\boldsymbol{\theta},\boldsymbol{z})\in\mathbb{R}^N\times\mathbb{R}^{N-1}}{\text{minimize}}\left\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|_2^2+\lambda\|\boldsymbol{z}\|_2\right\},\quad \text{其约束为}\boldsymbol{D\theta}=\boldsymbol{z} \tag{4.62}$$

在此引入辅助向量 $\boldsymbol{z}\in\mathbb{R}^{N-1}$，$\boldsymbol{D}$ 是一阶差分的 $(N-1)\times N$ 矩阵。现在来看问题的拉格朗日形式，即

$$L(\boldsymbol{\theta},\boldsymbol{z};\boldsymbol{u}):=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|_2^2+\lambda\|\boldsymbol{z}\|+\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{D\theta}-\boldsymbol{z}) \tag{4.63}$$

其中 $\boldsymbol{u}\in\mathbb{R}^{N-1}$ 是一个拉格朗日乘子向量。简单计算可得到拉格朗日对偶函数 $\mathcal{Q}$

$$\mathcal{Q}(\boldsymbol{u}):=\inf_{(\boldsymbol{\theta},\boldsymbol{z})\in\mathbb{R}^N\times\mathbb{R}^{N-1}}L(\boldsymbol{\theta},\boldsymbol{z};\boldsymbol{u})=\begin{cases}-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D}^{\mathrm{T}}\boldsymbol{u}\|^2, & \|\boldsymbol{u}\|_\infty\leqslant\lambda\\ -\infty, & \text{其他}\end{cases} \tag{4.64}$$

拉格朗日对偶问题要最大化 $\mathcal{Q}(\boldsymbol{u})$，由此得到最优解 $\widehat{\boldsymbol{u}}=\widehat{\boldsymbol{u}}(\lambda)$，可通过 $\widehat{\boldsymbol{\theta}}=\boldsymbol{y}-\boldsymbol{D}^{\mathrm{T}}\widehat{\boldsymbol{u}}$ 来求得原问题的最优解 $\widehat{\boldsymbol{\theta}}=\widehat{\boldsymbol{\theta}}(\lambda)$。对偶计算见习题 4.23。

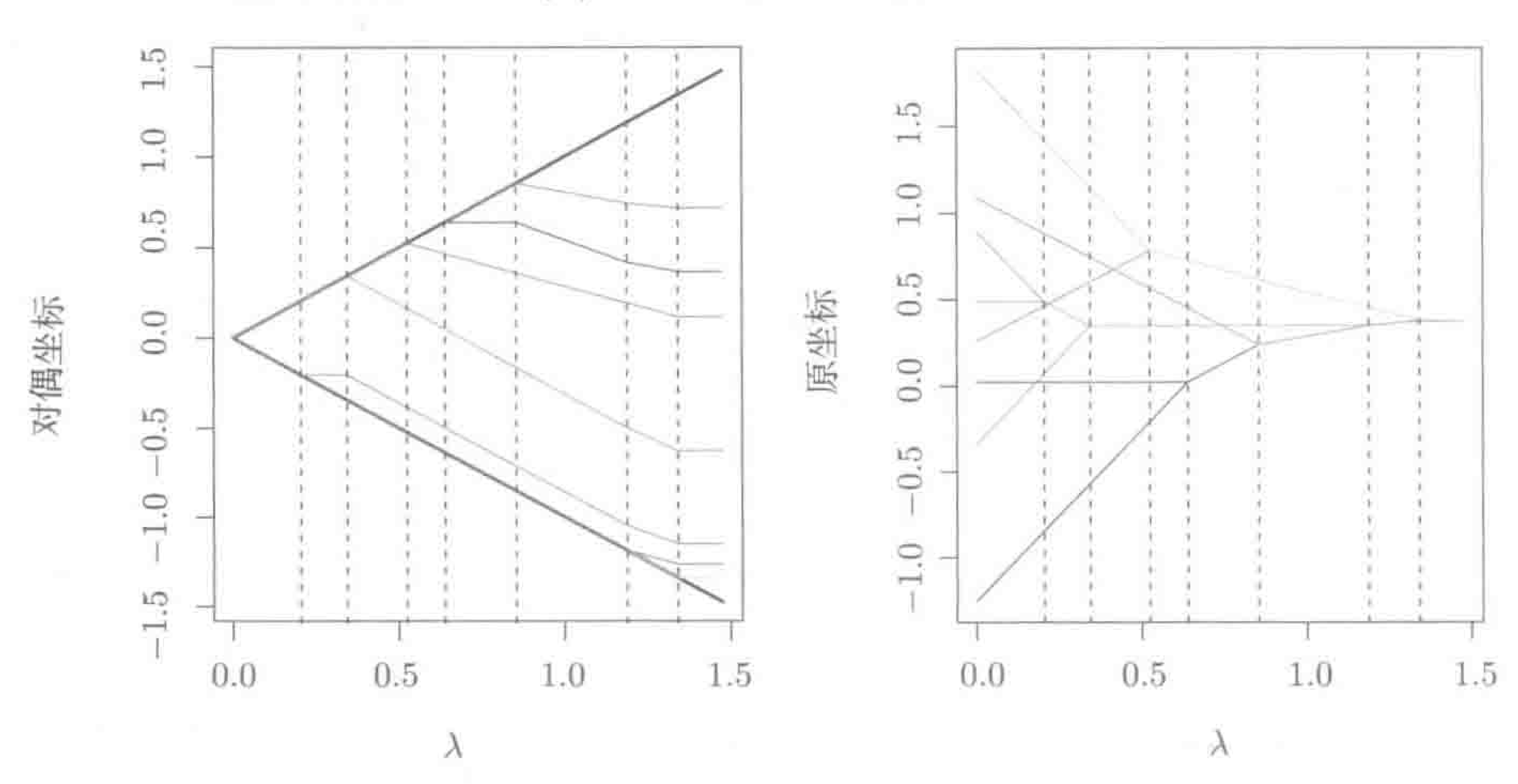


图 4-9　采用对偶路径算法的例子。左图为 $\widehat{\boldsymbol{u}}(\lambda)$，右图为 $\widehat{\boldsymbol{\theta}}(\lambda)$。可以看到在对偶坐标下，当参数接近边界时，在原始坐标下没有出现融合

当正则化参数 λ 足够大时，对偶就会最大化，或等价于最小化 $-\mathcal{Q}(\boldsymbol{u})$，这就可以转化为一个非约束的线性回归问题，其最优解为

$$\boldsymbol{u}^*:=\left(\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}\right)^{-1}\boldsymbol{D}\boldsymbol{y} \tag{4.65}$$

当 λ 降到关键水平 $\|\boldsymbol{u}^*\|_\infty$ 时，约束条件起作用。λ 减小时，一旦最优解的元素 $\widehat{u}_j(\lambda)$ 碰到边界 λ，它们就不会离开边界（Tibshirani$_2$ and Taylor 2011）。这个特性引出了一个十分直接的路径算法，类似于 5.6 节中 LARS 的思想。图 4-9 给出了对偶路径算法的应用示例，详见习题 4.23。

4. **动态规划算法在融合 lasso 上的运用**

动态规划是一种计算方法，将复杂问题分解为简单的子问题来求解。在一维融合 lasso 问题上，变量的线性排序意味着在固定任意变量时，可将问题分割为两个子问题，分别讨论固定变量左边和右边的情况。向前传递（forward pass）需从左向右移动，固定一个变量，将其左边的变量视为该固定变量的函数并求解。当求解到最右时，向后传递（backward pass）即可求得完整的解。

Johnson（2013）提出采用这种动态规划算法来求解融合 lasso。具体而言，要将式（4.59）中与 θ_1 有关的项分离出来，重写目标函数（4.59）为

$$f(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}(y_1-\theta_1)^2 + \lambda|\theta_2-\theta_1|}_{g(\theta_1,\theta_2)} + \left\{\frac{1}{2}\sum_{i=2}^{N}(y_i-\theta_i)^2 + \lambda\sum_{i=3}^{N}|\theta_i-\theta_{i-1}|\right\} \tag{4.66}$$

由此分解，得到了向前传递第一步所需求解的子问题：计算

$$\widehat{\theta}_1(\theta_2) := \arg\min_{\theta_1\in\mathbb{R}} g(\theta_1,\theta_2)$$

消除第一个变量。现在可通过

$$f_2(\theta_2,\ldots,\theta_N) = f(\widehat{\theta}_1(\theta_2),\theta_2,\ldots,\theta_N) \tag{4.67}$$

来求解简化后的目标函数 $f_2:\mathbb{R}^{N-1}\to\mathbb{R}$。然后在 θ_2 上重复这个过程，以此求得 $\widehat{\theta}_2(\theta_3)$，直至得到 $\widehat{\theta}_N$。接下来通过回代求得 $\widehat{\theta}_{N-1}=\widehat{\theta}_{N-1}(\widehat{\theta}_N)$，再依次求得 $\widehat{\theta}_{N-2},\ldots,\widehat{\theta}_2,\widehat{\theta}_1$。

如果每个参数 θ_i 只能取得 K 个不同值，则最小化的值 $\widehat{\theta}_j(\theta_{j+1})$ 很容易求解，并保存为一个 $K\times K$ 的矩阵。在连续情况下，需最小化的函数是分段线性的二次函数，需要一种有效的方式计算并存储相关信息，详见文献 Johnson（2013）。这是目前已知的最快算法，只需要 $\mathcal{O}(N)$ 次计算，比上面描述的路径算法要快很多。有趣的是，如果将基于 ℓ_1 差分惩罚改为基于 ℓ_0 的惩罚，这种方法依然能用，但问题不再是凸的。习题 4.24 要读者自己去实现这种离散情况。

4.5.2 趋势滤波

融合 lasso 中的一阶绝对值差分惩罚可推广到高阶差分惩罚，这会得到目标函数

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\theta_i)^2 + \lambda\cdot\left\|\boldsymbol{D}^{(k+1)}\theta\right\|_1\right\} \tag{4.68}$$

这就是趋势滤波（trend filtering）。这里 $\boldsymbol{D}^{(k+1)}$ 是一个 $(N-k-1)\times N$ 维矩阵，这个矩阵为 $k+1$ 阶离散差分。融合 lasso 使用一阶差分（$k=0$），而高阶差分则

促使高阶光滑。一般来说，k 阶趋势滤波得到的解是自由度为 k 的分段多项式。线性趋势滤波（$k=1$）尤其有用，它会得到分段线性解。解中的节点并不需要特别指明，会从凸优化过程中直接得到。针对这个问题，Kim, Koh, Boyd and Gorinevsky（2009）提出了一种有效的内部点算法。Tibshirani_2（2014）证明趋势滤波估计远比光滑样条更能适应光滑度的局部水平，这与局部自适应回归样条十分相似。另外他还证明，对于 k 阶导数为有界变差的函数，估计会以 minimax 速率收敛到真正的底层函数（光滑样条之类的线性估计没有这个性质）。此外，Tibshirani_2 and Taylor（2001）证明，带有 m 个节点的解的自由度估计为 $\text{df}=m+k+1$。[①]

图 4-10 为在分段线性函数采用趋势滤波法来拟合空气污染数据。为了进行比较，图中还包含了光滑样条拟合，其自由度为 $\text{df}=4$。虽然两者拟合相似，但趋势滤波能找到数据的自然拐点。

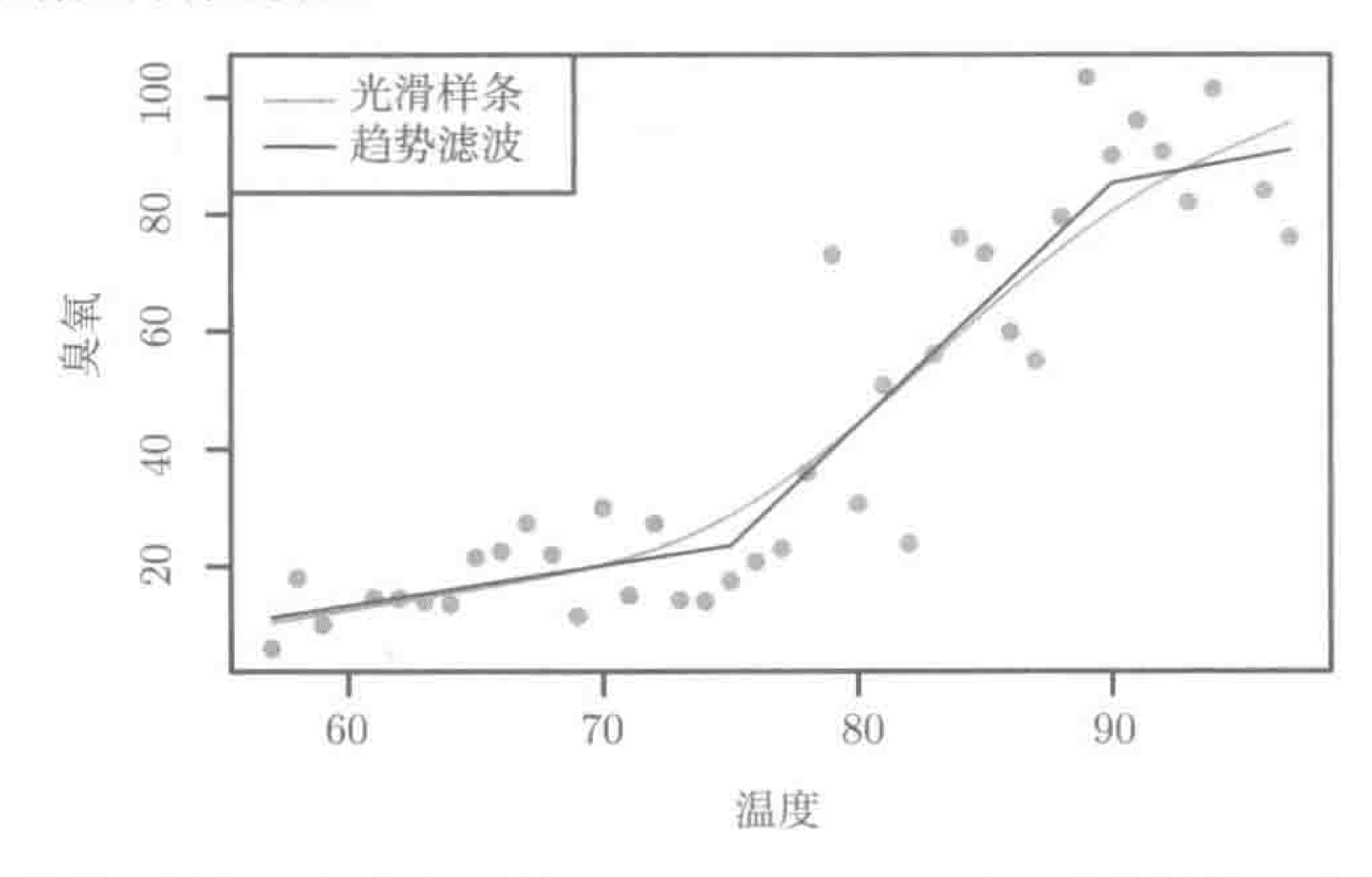


图 4-10　分段线性函数采用趋势滤波拟合空气污染数据。为了进行比较，图中还包含了同一自由度的光滑样条

在式（4.68）中，假设观测按等间距位置进行。为了适应任意（顺序）位置 x_i，惩罚项可修改为（Tibshirani_2 2014）

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\theta_i)^2+\lambda\cdot\sum_{i=1}^{N-2}\left|\frac{\theta_{i+2}-\theta_{i+1}}{x_{i+2}-x_{i+1}}-\frac{\theta_{i+1}-\theta_i}{x_{i+1}-x_i}\right|\right\} \tag{4.69}$$

这里比较了邻接对的经验斜率（empirical slope），并使二者相同。这是图 4-10 用到的惩罚，因为温度值并不均匀分布。

4.5.3　近保序回归

Tibshirani_2, Hoefling and Tibshirani（2011）提出，可以对一维融合 lasso 进行简单的改进，从而使解单调。这一点的基础是对保序回归（isotonic regression）的

① 这是自由度的一个无偏估计，见 2.5 节。

松弛。保序回归的经典形式是通过最小化带约束的问题

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\text{minimize}}\left\{\sum_{i=1}^{N}(y_i-\theta_i)^2\right\},\quad \text{其约束为}\theta_1\leqslant\theta_2\leqslant\ldots\leqslant\theta_N \tag{4.70}$$

来估计 $\boldsymbol{\theta}\in\mathbb{R}^N$。由此而来的解是对数据的最佳单调（非增）拟合。单调非递增的解可以通过先反转数据符号而得到。问题（4.70）有唯一解，可以通过 PAVA（Pool Adjacent Violators Algorithm）求得（Barlow, Bartholomew, Bremner and Brunk 1972）。

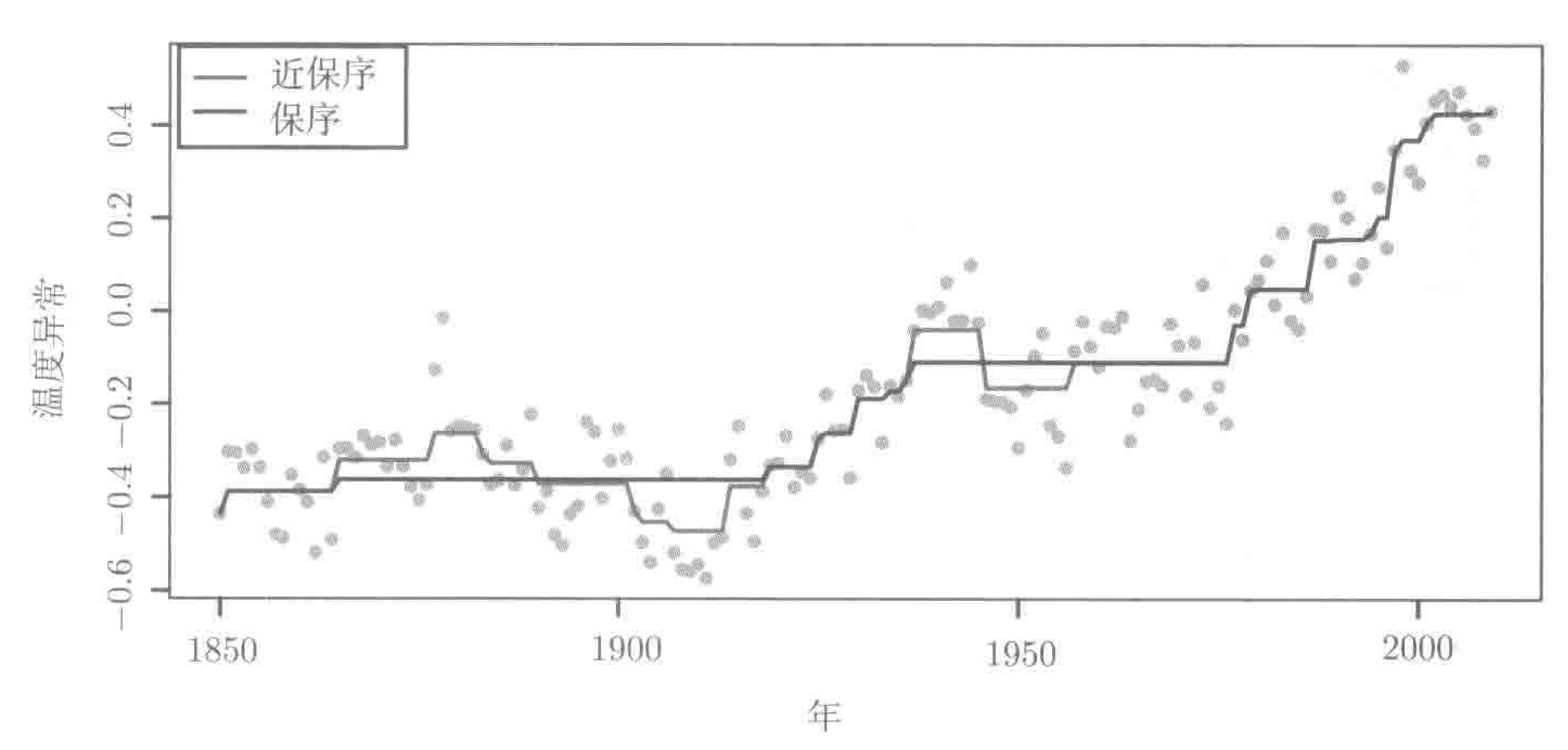


图 4-11 用近保序回归拟合全球变暖数据，图中给出了每年的温度异常。通过交叉验证选取 λ 值，拟合结果能支持数据中的非单调特性

近保序回归是一种自然的松弛，这里引入一个正则化参数 $\lambda\geqslant 0$，并求解优化问题

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\theta_i)^2+\lambda\sum_{i=1}^{N-1}(\theta_i-\theta_{i+1})_+\right\} \tag{4.71}$$

惩罚项用于惩罚违反单调特性的邻接对，即 $\theta_i>\theta_{i+1}$。当 $\lambda=0$ 时，其解来自对数据进行的插值处理，当 $\lambda\to\infty$，需重新求得经典保序回归问题（4.70）的解。λ 的中间值得到的是非单调解，需要在单调性与拟合优度之间进行权衡。这种权衡可以用于评估给定数据序列单调假设的有效性。图 4-11 对 1856~1999 每年温度异常数据采用该方法，这些数据相对于 1961~1990 的均值。近保序回归问题（4.71）的解可以通过对路径算法进行简单修改而得到，这个过程类似于式（4.70）用到的 PAVA 算法；详见文献 Tibshirani$_2$ et al.（2011）。

4.6 非凸惩罚

通过将 ℓ_2 惩罚修改为 ℓ_1 惩罚，可以看出对同样的**有效 df**，lasso 会选择一个系数非零的变量子集，并将这些变量的系数收缩得较小。当 p 很大而相关变量的数目比较小时，这可能还不够。为了充分减小所选择的变量子集，lasso 可能结束过度收缩以保留变量。正因如此，非凸惩罚一直受到关注。

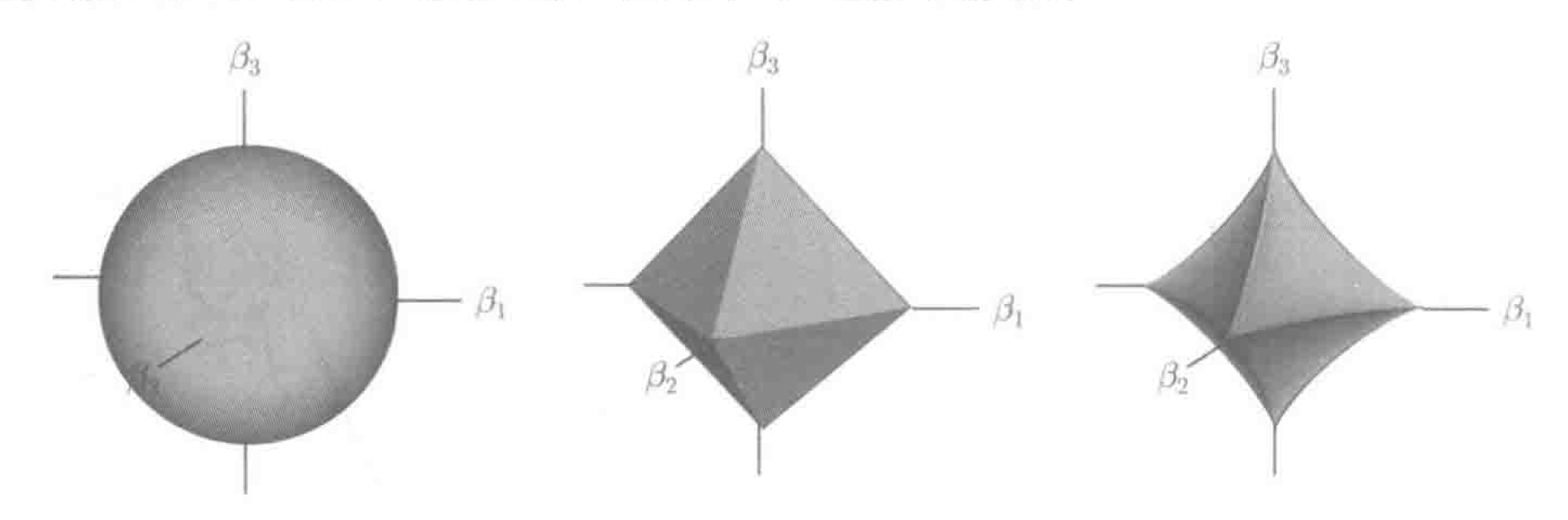


图 4-12　$\mathbb{R}^3$ 空间中的 ℓ_q 单位球，$q=2$（左），$q=1$（中），$q=0.8$（右）。$q<1$ 时，约束区域是非凸的。小一点的 q 对应着小一些的非零系数，小一些变量压缩。非凸性带来了组合优化难题

通常可以选择 ℓ_q 惩罚（$0 \leqslant q \leqslant 1$），极端情况采用 ℓ_0 会选择最优子集。图 4-12 比较了 $q \in \{2,1,0.8\}$ 下的 ℓ_q 单位球。图中最右边的"球"像尖钉一样，它的非凸特性意味着在这种约束条件下将选择边和坐标轴。然而，非凸特性带来了复杂的组合计算问题。即使在最简单的 ℓ_0 情况下，也只有在 $p \approx 40$ 或者更小时才能求解。因为这个原因和相关的统计原因，其他非凸惩罚方法被提出。这些方法包括 SCAD（Fan and Li 2001, *smoothly clipped absolute deviation*）和 MC+（Zhang 2010, *minimax concave*）惩罚。图 4-13 为 $\mathbb{R}^1$ 空间上 MC+ 惩罚族中的四种情形，按照非凸性参数 $\gamma \in (1, \infty)$ 排序。它们填补了 lasso（$\gamma=\infty$）和最优子集（$\gamma=1_+$）之间的空白。惩罚函数为分段的二次函数（见习题 4.25），重要的是响应阈值函数是分段线性且连续的。具体而言，采用平方误差损失，会得到（非凸）的优化问题

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{minimize}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \sum_{j=1}^{p} P_{\lambda,\gamma}(\beta_j) \right\} \tag{4.72}$$

定义各个坐标上的 MC+ 惩罚为

$$P_{\lambda,\gamma}(\theta) := \int_0^{|\theta|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ \mathrm{d}x \tag{4.73}$$

通过坐标下降法来求解式（4.72）的一维版（标准形式）

$$\underset{\beta\in\mathbb{R}^1}{\text{minimize}}\left\{\frac{1}{2}\left(\beta-\widetilde{\beta}\right)^2+\lambda\int_0^{|\beta|}\left(1-\frac{x}{\lambda\gamma}\right)_+\mathrm{d}x\right\} \tag{4.74}$$

对 $\gamma>1$，有单调[①]唯一解

$$\mathcal{S}_{\lambda,\gamma}\left(\widetilde{\beta}\right)=\begin{cases}0, & \left|\widetilde{\beta}\right|\leqslant\lambda\\ \operatorname{sgn}(\widetilde{\beta})\left(\dfrac{\left|\widetilde{\beta}\right|-\lambda}{1-\dfrac{1}{\gamma}}\right), & \lambda<\left|\widetilde{\beta}\right|\leqslant\lambda\gamma\\ \widetilde{\beta}, & \left|\widetilde{\beta}\right|>\lambda\gamma\end{cases} \tag{4.75}$$

图 4-13 中的右图为式（4.75）的示例。大的 $\widetilde{\beta}$ 值留下，小的值设置为零，收缩中间值。随着 γ 减小，中间区域变窄，直到最终变为最优子集（图中橘色线）的硬阈值函数。通过比较会发现，ℓ_q 族（$q<1$）阈值函数对于 $\widetilde{\beta}$ 是非连续的。

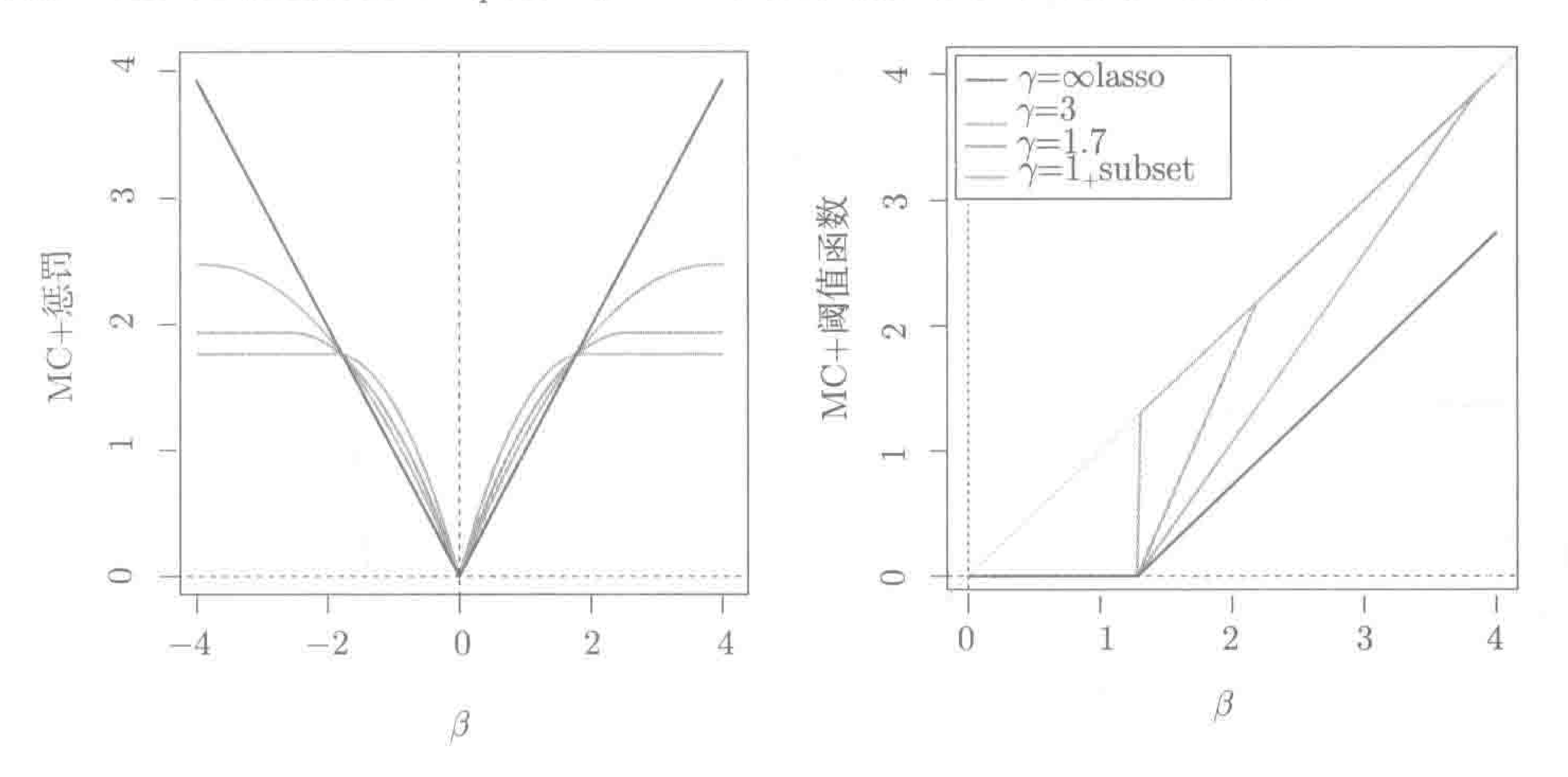


图 4-13 左图：非凸稀疏惩罚 MC+ 族，按照稀疏性参数 $\gamma\in(1,\infty)$ 排序。右图：MC+ 分段连续阈值函数（只画出了第一象限），这样的惩罚族适合坐标下降算法

Mazumder, Friedman and Hastie（2011）对整个 MC+ 族求拟合解的路径时采用了坐标下降法，发现了 $\mathcal{S}_{\lambda,\gamma}$（在 λ 和 γ 上）的连续性。他们开发的 R 语言包 sparsenet（Mazumder, Hastie and Friedman 2012）从 lasso 解开始，沿着 γ 序列移向更稀疏的模型，对每个模型用 λ 拟合一个正则化路径。尽管这并不能说解决了非凸问题（4.72），但这个方法能很快找到好的解。

① 函数非凸，但其在 $\mathbb{R}^1$ 空间上仍有单调解；p 维问题（4.72）则不一定如此。

Zou（2006）提出了**自适应 lasso 法**，它作为拟合模型的方法比 lasso 更加稀疏。采用初步估计 $\widetilde{\beta}$，自适应 lasso 需要求解目标函数

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\} \tag{4.76}$$

其中 $w_j = 1/|\widetilde{\beta}_j|^\nu$。自适应 lasso 惩罚可以看作是 ℓ_q 惩罚（其中 $q = 1 - \nu$）的一个近似。自适应 lasso 的一个优点是，当给定初步估计后，式（4.76）对 β 是凸的。此外，如果初步估计是 $\sqrt{N}$ 一致的，Zou（2006）已证明，在更一般的条件下，该方法比 lasso 更易恢复真正的模型。如果 $p < N$，可以采用最小二乘解作为初步估计；如果 $p \geqslant N$，最小二乘估计无法使用，但是单变量回归系数可以用来作为初步估计，在一定的条件下能得到好的复原特性（Huang, Ma and Zhang 2008）。习题 4.26 会研究自适应 lasso 与 2.8 节非负 garrote 之间的联系。

在本节最后介绍一下其他实际用于建立稀疏模型路径的非凸优化方法。逐步前向法（Hastie et al. 2009，Chapter 3）十分有效，在寻找好的稀疏子集方面非常不错。逐步前向法是一种贪心算法，即在各个步骤固定模型中已经存在的变量值，然后在剩余变量中选出最好的。逐步前向法的模型路径的理论性质不易理解，这部分归咎于算法过程的定义，这一点与优化问题的解相反。

参考文献注释

弹性网方法是 Zou and Hastie（2005）提出的，其原始的版本（和这里介绍的相似）和去偏的版本（试图取消岭收缩中的有偏影响）有所不同。Friedman et al.（2015）针对拟和弹性网惩罚的广义线性模型采用坐标下降算法求解，这已经在 R 的 glmnet 包中实现了。Yuan and Lin（2006a）引入了组 lasso，他们发表的论文引发了相关研究的热潮。Meier et al.（2008）将组 lasso 扩展到逻辑斯蒂回归问题中，而 Zhao, Rocha and Yu（2009）给出了一种更广义的结构化惩罚族，在这种情形下，组 lasso 只是一个特例。一系列的理论研究工作都试图给出组 lasso 估计何时比普通 lasso 统计偏差更低。Huang and Zhang（2010）以及 Lounici, Pontil, Tsybakov and van de Geer（2009）确定了组 lasso 的误差边界，可以看出在一些特定情况下组 lasso 比普通 lasso 表现要好。Negahban, Ravikumar, Wainwright and Yu（2012）提出了一个 M 估计分析的一般性框架，其中组 lasso 是一个特例，而且还有其他更广义的结构性惩罚。Obozinski, Wainwright and Jordan（2011）关注多变量回归下组 lasso 是否比普通 lasso 拥有更好的变量选择性能。

重叠组 lasso 由 Jacob et al.（2009）引入，稀疏组 lasso 由 Puig, Wiesel and Hero（2009）以及 Simon, Friedman, Hastie and Tibshirani（2013）引入。已经有很

多种算法用于求解组 lasso、重叠组 lasso，而且还有很多结构化的推广，详见 Bach, Jenatton, Mairal and Obozinski（2012）给出的总结。

加法模型由 Stone（1985）提出，它可以看作解决非参数回归中维数灾难的一种方法；Hastie and Tibshirani（1990）给出了（广义）加法模型的更多背景。COSSO 模型是 Lin and Zhang（2003）在再生核希尔伯特空间和 ANOVA 样条分解情境下提出的。Wahba（1990）和 Gu（2002）的书中提供了关于样条和 RKHS 的更多背景。Ravikumar et al.（2009）紧跟着提出了 SPAM 模型，这看起来更简单且更一般，而且确立了估计量高维一致性的某种形式。Meier, van de Geer and Buhlmann（2009）基于经验 L^2 范数的显示惩罚，研究了一个相关估计族，对应 $\lambda_{\mathcal{H}}=0$ 的双倍惩罚估计。Koltchinski and Yuan（2008, 2010）分析了 COSSO 估计和双倍惩罚估计（4.52）。Raskutti et al.（2009, 2012）推导出了稀疏加法模型的极大极小边界，证明了双倍惩罚估计（4.52）能对多种 RKHS 族达到这些边界，包括样条函数这样的特例。

融合 lasso 是由 Tibshirani, Saunders, Rosset, Zhu and Knight（2005）提出的。现在不同版本的融合 lasso 已经有多种求解算法，包括 Hoefling（2010）、Johnson（2013）和 Tibshirani_2 and Taylor（2011）提出的算法。

MC+ 阈值函数首先由 Gao and Bruce（1997）在小波收缩的情况下提出。对稀疏模型来说，非凸惩罚有很多优势。Zou and Li（2008）发明了局部线性近似算法，来求解非凸优化问题。Mazumder et al.（2011）中也讨论了这类算法和其他算法。

习　题

习题 4.1　假设有两个相同变量 $X_1=X_2$，一个响应变量 Y，并执行有惩罚参数 $\lambda>0$ 的岭回归 [见 2.2 节的式（2.7）]。求系数估计 $\widehat{\beta}_j(\lambda)$。

习题 4.2　4.2 节中带有噪声的两个相同示例中，有两个变量强正相关。画出损失函数和惩罚函数的轮廓图，证明为什么弹性网方法相对于 lasso 法更能实现系数共享。

习题 4.3　对于弹性网问题（4.2），求解

(a) 如何通过归一化各个预测子来简化 $\widehat{\beta}_0$ 的计算，使得对于任意 λ，有 $\widehat{\beta}_0=\bar{y}$。如何再反向得到未归一化预测子下的 $\widehat{\beta}_0$ 估计？

(b) 对于 $\widehat{\beta}_j$ 的更新，验证基于坐标下降法的软阈值表达式（4.4）。

习题 4.4　如果一些变量是因子，组 lasso 问题（4.5）的解会有什么不同？求证：当模型中有截距时，各个因子的最优系数之和为零。

习题 4.5　该习题研究组 lasso 中惩罚项变化的影响。考虑组 lasso（4.3.1 节）中的约

束条件 $\|\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j\|_2 < \lambda$，假设 $\boldsymbol{r}_j$ 是独立同分布的噪声，均值为 $\mathbf{0}$，方差为 $\sigma^2\boldsymbol{I}$。

求证：

$$\mathbb{E}\left\|\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j\right\|_2^2 = \sigma^2\left\|\boldsymbol{Z}\right\|_F^2 \tag{4.77}$$

为了公平比较组 lasso 中的惩罚项，应该将式（4.5）中的 $\lambda\sum_{j=1}^{J}\|\theta_j\|_2$ 替换为

$$\lambda\sum_{j=1}^{J}\tau_j\left\|\theta_j\right\|_2 \tag{4.78}$$

其中 $\tau_j = \|\boldsymbol{Z}_j\|_F$。求证当 $\boldsymbol{Z}_j$ 为标准正交时，会有 $\tau_j = \sqrt{p_j}$。

习题 4.6　证明在标准正交条件 $\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j = \boldsymbol{I}$ 下，更新式（4.15）可通过求解不动点式（4.13）得到。

习题 4.7　考虑组 lasso 的块状解向量式（4.14）。如果已知 $\|\widehat{\theta}_j\|$，就能够求出闭合解。设 $\boldsymbol{Z}_j$ 的奇异值分解为 $\boldsymbol{Z}_j = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$。设 $r^* = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{r}_j \in \mathbb{R}^{p_j}$。证明 $\phi = \|\widehat{\theta}_j\|$ 时，等式

$$\sum_{\ell=1}^{p_j}\frac{r_\ell^{*2}d_\ell^2}{\left(d_\ell^2\phi+\lambda\right)^2} = 1 \tag{4.79}$$

成立，其中 d_ℓ 是 $\boldsymbol{D}$ 的第 ℓ 个对角元素。给出如何用黄金搜索算法求解 ϕ。编写一个 R 函数实现这个搜索算法。

习题 4.8　讨论哑变量矩阵（哑变量表示因子，有 p_j 为层级）的归一化 $\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j = \boldsymbol{I}$ 的影响。采用对照而非哑变量是否可以缓解这种情况？

习题 4.9　采用 5.3.3 节的方法，给出组 lasso 问题的广义梯度更新式（4.16a）。写一个 R 函数（对于单个组）实现这个算法，该算法应有一个选项可以用来指定是否要采用 Nesterov 加速。

习题 4.10　如何采用 5.3.3 节的方法得出稀疏组 lasso 问题的广义梯度更新式（4.23）？

习题 4.11　增加维数，比较习题 4.7 和习题 4.9 中算法的性能。确定这两种算法产生的解是一样的。比较它们的计算速度，比如，在 R 中可以用命令`system.time()`得到算法的执行时间。

习题 4.12　在稀疏组 lasso 中，条件式（4.19）可让 $\widehat{\theta}_j$ 为零，定义

$$\begin{aligned}J(t) &= \frac{1}{\lambda(1-\alpha)}\left\|\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j - \lambda\alpha\cdot t\right\|_2 \\ &= \|s\|_2\end{aligned} \tag{4.80}$$

现在求解

$$\min_{t:t_k\in[-1,1]} J(t) \tag{4.81}$$

求证：当且仅当 $\|\widehat{g}_j\|_2 \leqslant \lambda(1-\alpha)$ 且 $\widehat{g}_2 = \mathcal{S}_{\lambda\alpha}\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{r}_j\right)$ 时，$\widehat{\theta}_j = 0$。

习题 4.13　如果 $\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j = \boldsymbol{I}$，求证：式（4.21）可由式（4.12）得到。

习题 4.14　考虑例 4.3 中的分层相互作用公式和优化问题（4.29）～（4.31）。

(a) 为何乘子 p_1 和 p_2 在第三个惩罚中有意义？

(b) 假设式 (4.29) 中第三个矩阵增加一个元素全为 1 的向量 $[\mathbf{1}\ \boldsymbol{Z}_1\ \boldsymbol{Z}_2\ \boldsymbol{Z}_{1:2}]$，并且增大参数向量 $\widetilde{\mu}$。现在替换第三个组惩罚项为

$$\sqrt{p_1 p_2 \widetilde{\mu}^2 + p_2 \|\widetilde{\alpha}_1\|_2^2 + p_1 \|\widetilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2}$$

求证对于任何 $\lambda > 0$，$\widehat{\widetilde{\mu}} = 0$。

(c) 对于任意 $\lambda > 0$，式（4.29）～（4.31）的解等价于式（4.32）的解。如何将后面式子的解映射成前面的解？

习题 4.15　对于稀疏加法模型

$$\begin{aligned} &\underset{\beta\in\mathbb{R}^J,\{f_j\in\mathcal{F}_j\}_1^J}{\operatorname{minimize}} \quad \mathbb{E}\left(Y - \sum_{j=1}^{J}\beta_j f_j(X_j)\right)^2 \\ &\text{其约束为} \quad \|f_j\|_2 = 1\forall j \\ &\qquad\qquad\quad \sum_{j=1}^{J}|\beta_j| \leqslant t \end{aligned} \tag{4.82}$$

这个公式不是凸的，但对于 β 和 $\{f_j\}_1^J$ 双凸（biconvex）。求证：可将 β_j 吸收进 f_j，求解式（4.82）等价于求解凸的式（4.38）：

$$\underset{f_j\in\mathcal{F}_j, j=1,\ldots,J}{\operatorname{minimize}}\left\{\mathbb{E}\left[Y - \sum_{j=1}^{J} f_j(X_j)\right]^2 + \lambda\sum_{j=1}^{J}\|f_j\|_2\right\}$$

（Ravikumar et al. 2009）。

习题 4.16　SPAM backfitting 方程（4.40）是属于函数更新，其中 $\widehat{f}_j$ 是拟合后的函数，由光滑算子 $\mathcal{S}_j$ 得到，N 向量形式 $\boldsymbol{f}_j$ 是 f_j 由 X_j 的 N 样本值得到。假设光滑算子 $\mathcal{S}_j$ 拟合下面这种的线性展开式

$$f_j(\cdot) = \sum_{\ell=1}^{p_j}\beta_{j\ell}\psi_{j\ell}(\cdot) \tag{4.83}$$

其中 $\beta_j = [\beta_{j1}\quad \beta_{j2}\quad \ldots\quad \beta_{jp_j}]$ 是系数向量。

(a) 假设基矩阵是正交的：$\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\psi}_j = \boldsymbol{I}_{p_j}$。求证：对于参数 θ_j，SPAM backfitting 方程等价于普通组 lasso 估计方程。

(b) 如果 $\boldsymbol{\psi}_j$ 不是正交的，又会如何？

习题 4.17　求证 COSSO 问题（4.44）的任意最优解都属于 $\mathcal{H}_0$，$\mathcal{H}_0$ 是由核函数 $\{\mathcal{R}(\cdot, x_i), i = 1, \ldots, N\}$ 组成的希尔伯特空间。这里会用到一个公理：任意函数 $f \in \mathcal{H}$ 都可以分解为 $g + h$ 的形式，其中 $g \in \mathcal{H}_0$，h 与 $\mathcal{H}_0$ 正交，这意味着对所有的 $f_0 \in \mathcal{H}_0$，有 $\langle h, f_0\rangle_{\mathcal{H}}$。

(a) 对上面的函数 $f = g + h$，求证 $\frac{1}{N}\sum_{i=1}^{N}(y - f(x_i))^2$ 只与 g 有关。(提示：这里会用到再生核的性质。)

(b) 求证惩罚项由于包含了 $h \neq 0$ 的分量，只会递增。由此得出，所有的最优解 $\widehat{f}$ 都属于 $\mathcal{H}_0$。

习题 4.18　当 $\lambda = \tau^4/4$ 时，验证式（4.47）中 f_j 的解与式（4.44）的解相一致。

习题 4.19　对于加法模型式（4.42），假设每个函数 f_j 与一个再生核函数 $\mathcal{R}_j$ 相关，由此得到目标函数

$$\underset{\substack{\theta_j \in \mathbb{R}^N \\ j=1,\ldots,J}}{\operatorname{minimize}} \left\{ \left\| \boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{R}_j \boldsymbol{\theta}_j \right\|^2 + \lambda \sum_{j=1}^{J} \frac{1}{\gamma_j} \boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{R}_j \boldsymbol{\theta}_j \right\} \tag{4.84}$$

（$1/N$ 已经包含在 λ 中）。

(a) 定义 $\widetilde{\boldsymbol{R}}_j = \gamma_j \boldsymbol{R}_j$，$\widetilde{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j/\gamma_j$。在新的参数下，求证 $\widetilde{\boldsymbol{\theta}}_j$ 的估计方程为

$$-\widetilde{\boldsymbol{R}}_j (\boldsymbol{y} - \boldsymbol{f}_+) + \lambda \widetilde{\boldsymbol{R}}_j \widetilde{\boldsymbol{\theta}}_j = \boldsymbol{0},\ j = 1, \ldots, J \tag{4.85}$$

其中 $\boldsymbol{f}_+ = \sum_{j=1}^{J} \boldsymbol{f}_j$，$\boldsymbol{f}_j = \widetilde{\boldsymbol{R}}_j \widetilde{\boldsymbol{\theta}}_j$。

(b) 证明上面的式子可以重写为

$$\widetilde{\boldsymbol{\theta}}_j = \left(\widetilde{\boldsymbol{R}}_j + \lambda \boldsymbol{I}\right)^{-1} r_j \tag{4.86a}$$

$$\widetilde{\boldsymbol{f}}_j = \widetilde{\boldsymbol{R}}_j \left(\widetilde{\boldsymbol{R}}_j + \lambda \boldsymbol{I}\right)^{-1} r_j \tag{4.86b}$$

其中 $\boldsymbol{r}_j = \boldsymbol{y} - \boldsymbol{f}_+ + \boldsymbol{f}_j$。

(c) 定义 $\widetilde{\boldsymbol{R}}_+ = \sum_{j=1}^{J} \widetilde{\boldsymbol{R}}_j = \sum_{j=1}^{J} \gamma_j \boldsymbol{R}_j$。求证：

$$\boldsymbol{f}_+ = \widetilde{\boldsymbol{R}}_+ \left(\widetilde{\boldsymbol{R}}_+ + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{y} = \widetilde{\boldsymbol{R}}_+ \boldsymbol{c} \tag{4.87a}$$

$$\boldsymbol{c} = \left(\widetilde{\boldsymbol{R}}_+ + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{y} \tag{4.87b}$$

并与之前项进行比较。

(d) 求证 $\widetilde{\boldsymbol{\theta}}_j = \boldsymbol{c} \forall j$。所以即使在式子中有 J 个 N 维参数 $\widetilde{\boldsymbol{\theta}}_j$，它们的估计也都是一样。

这表明在给定 $\gamma_j \boldsymbol{f}_j = \gamma_j \boldsymbol{R}_j \boldsymbol{c} = \gamma_j \boldsymbol{g}_j$ 时，证明在用其他方法拟合 COSSO 模型时，第二步（4.46）存在合理性（见 4.4 节）。

习题 4.20　证明双重正则化估计（4.52）的任意最优解都有 $\widehat{f}_j(\cdot)=\sum_{i=1}^{N}\widehat{\theta}_{ij}\mathcal{R}(\cdot,x_{ij})$，其中权重 $(\widehat{\theta}_j, j=1,\ldots,J)$ 是通过求解凸问题（4.53）得到的。

习题 4.21　考虑融合 lasso 问题（4.56），描述 $\widehat{\beta}_0$ 的性质。求证如果通过减去样本均值来归一化预测变量和响应变量，则可以忽略 β_0 项，估计值 $\widehat{\beta}_j$ 不受影响。现在考虑有常参数 θ_0 的融合 lasso 信号估计（4.54）的一个变体

$$\underset{\theta_0,\boldsymbol{\theta}}{\text{minimize}}\sum_{i=1}^{N}(y_i-\theta_0-\theta_i)^2+\lambda_1\sum_{i=1}^{N}|\theta_i|+\lambda_2\sum_{i=2}^{N}|\theta_i-\theta_{i-1}| \tag{4.88}$$

求解 $\widehat{\theta}_0$，求证 $(\widehat{\theta}_i)$ 的中值为 0。

习题 4.22　对于线性变换（4.60）的矩阵 $\boldsymbol{M}$，解答

(a) 证明其逆矩阵 $\boldsymbol{M}^{-1}$ 是个下三角矩阵，对角线之下所有元素全为 1。

(b) 对于图 4-8 中的 CGH 数据，计算该矩阵的列之间的相关性。

(c) 运用 glmnet（`maxdef=200, type="naive"`）拟合模型（4.61），证明拟合值对应所研究的参数。比较同样情况下`lars`的性能。采用软阈值后处理算法来与图 4-8 进行比较。

习题 4.23　推导对偶优化问题（4.64）。假设 $\widehat{\boldsymbol{u}}(\lambda)$ 的第 k 个元素在 $\lambda=\lambda_k$ 边界，设集合 B 为它们的索引，$\boldsymbol{s}$ 是符号值向量。求证式（4.64）在 λ_k 处的解可通过求解目标函数

$$\underset{\boldsymbol{u}_{-B}}{\text{minimize}}\ \frac{1}{2}\left\|\boldsymbol{y}-\lambda_k\boldsymbol{D}_B^{\mathrm{T}}\boldsymbol{s}-\boldsymbol{D}_{-B}^{\mathrm{T}}\boldsymbol{u}_{-B}\right\|^2 \tag{4.89}$$

得到，解为 $\widetilde{\boldsymbol{u}}_B(\lambda)=\lambda\boldsymbol{s}$，$\widehat{\boldsymbol{u}}_{-B}(\lambda)=\left(\boldsymbol{D}_{-B}\boldsymbol{D}_{-B}^{\mathrm{T}}\right)^{-1}\boldsymbol{D}_{-B}\left(\boldsymbol{y}-\lambda\boldsymbol{D}_B^{\mathrm{T}}\boldsymbol{s}\right)$，$\lambda=\lambda_k$。$\widehat{\boldsymbol{u}}_{-B}(\lambda)$ 的各个元素的绝对值都小于 λ_k。求证解在 $\lambda<\lambda_k$ 时是分段线性的，保留解，直到 $\widehat{\boldsymbol{u}}_{-B}(\lambda)$ 的下一个元素达到边界值。求解可得到产生分段线性解时相应的元素和 λ 的值。

习题 4.24　这里采用动态规划来拟合融合 lasso。

(a) 在简单的情形中，每个 β_i 能取 K 个离散值中的一个，采用动态规划方法来拟合融合 lasso。

(b) 情形与 (a) 一样，用 ℓ_0 差分惩罚替代 ℓ_1 差分惩罚。比较两者在 CGH 数据上的表现。

习题 4.25　推导 4.6 节中一维 MC+ 准则（4.74）中的阈值函数（4.75）。

习题 4.26　求证在式（4.76）中，若 $\nu=1$，则自适应 lasso 的解与非负 garrote 的解（2.19）相似。特别地，如果约束自适应 lasso 的解与预估计一样有相同的符号，若选择一个合适的正则化参数，则它们和 garrote（2.19）有相同的解。

第5章　优化方法

5.1　引　　言

本章针对凸问题介绍一些基本的优化概念和算法，重点介绍与正则化估计（比如 lasso）相关的优化算法。这里主要介绍一阶算法，因为这类算法对于大规模的优化问题特别实用。本章会首先针对凸规划介绍一些基本的最优理论，然后考虑各种迭代算法。虽然这里只将重点放在凸问题上，但在本章后面也会涉及一些双凸问题。

5.2　凸优化条件

带凸约束条件的凸目标函数是一类重要的优化问题。对于集合 $\mathcal{C} \subseteq \mathbb{R}^p$，若 $\beta, \beta' \in \mathcal{C}$，且对任意的标量 $s \in [0,1]$，使得 $\beta(s) = s\beta + (1-s)\beta'$ 也属于集合 $\mathcal{C}$，则 $\mathcal{C}$ 为凸集。对于函数 $f : \mathbb{R}^p \to \mathbb{R}$，若对于 f 定义域的两个向量 β, β' 和任意的标量 $s \in (0,1)$，都有下式成立，则 f 为凸函数：

$$f(\beta(s)) = f(s\beta + (1-s)\beta') \leqslant sf(\beta) + (1-s)f(\beta') \tag{5.1}$$

这个不等式的几何意义为：连接 $f(\beta)$ 与 $f(\beta)$ 之间的弦总在 f 的图的上方，如图 5-1a 所示。这个不等式保证凸函数不会有不是全局最小值的局部最小值，如图 5-1b 所示。

5.2.1　优化可微问题

考虑优化问题

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}}\, f(\beta), \quad \beta \in \mathcal{C} \tag{5.2}$$

其中要最小化的 $f : \mathbb{R}^p \to \mathbb{R}$ 是凸目标函数，$\mathcal{C} \in \mathbb{R}^p$ 是凸的约束集。当目标函数 f 可微，则有全局最优解 $\beta^* \in \mathcal{C}$ 的充分必要条件为

$$\langle \nabla f(\beta^*), \beta - \beta^* \rangle \geqslant 0 \tag{5.3}$$

其中 $\beta \in \mathcal{C}$。充分条件很容易得到：对于任意的 $\beta \in \mathcal{C}$，有

$$f(\beta) \overset{\text{(i)}}{\geqslant} f(\beta^*) + \langle \nabla f(\beta^*), \beta - \beta^* \rangle \overset{\text{(ii)}}{\geqslant} f(\beta^*) \tag{5.4}$$

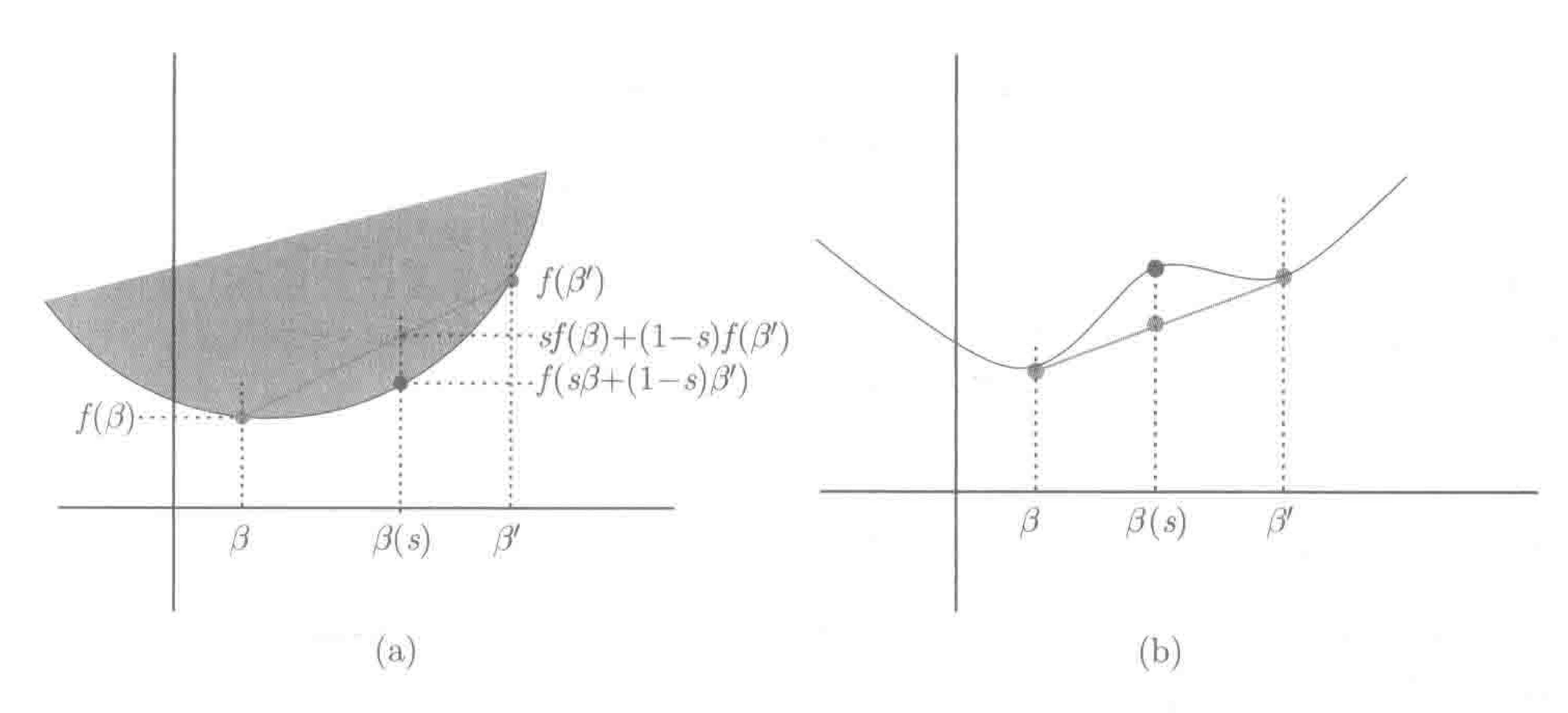


图 5-1 (a) 对于一个凸函数，直线 $sf(\beta)+(1-s)f(\beta')$ 总是在函数值 $f(s\beta+(1-s)\beta')$ 的上方；(b) 非凸函数会出现与不等式（5.1）不一样的情况。没有凸性，有可能得不到全局最小值，得到的只是局部最小值 β'

其中不等式 (i) 成立是因为 f 是凸函数 ①，不等式 (ii) 成立是因为优化条件式（5.3）。有一种特殊情况：当 $\mathcal{C}=\mathbb{R}^p$ 时，问题（5.2）成了无约束优化。一阶条件式（5.3）就简化成了经典的梯度为零的条件：$\nabla f(\beta^*)=0$。

通常，约束集 $\mathcal{C}$ 可以用凸约束函数的下水平集来描述。对于任意的凸函数 $g:\mathbb{R}^p\to\mathbb{R}$，由式（5.1）中的定义可知，下水平集 $\{\beta\in\mathbb{R}^p|g(\beta)\leqslant 0\}$ 为凸集。因此，凸优化问题可写成

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\, f(\beta),\quad \text{其约束为}\quad g_j(\beta)\leqslant 0,\quad j=1,\ldots,m \tag{5.5}$$

其中 $g_j(j=1,\ldots,m)$ 为满足约束的凸函数，这是式（5.2）的一种特殊形式。令 f^* 表示优化问题（5.5）的最优解的值。

与问题（5.5）相关的拉格朗日函数 $L:\mathbb{R}^p\times\mathbb{R}_+^m\to\mathbb{R}$ 为

$$L(\beta;\lambda)=f(\beta)+\sum_{j=1}^{m}\lambda_j g_j(\beta) \tag{5.6}$$

非负权重 $\lambda\geqslant 0$ 称为拉格朗日乘子。只要违反约束 $g_j(\beta)_j\leqslant 0$，乘子 λ_j 会起到惩罚的作用。事实上，若通过优化方法来得到 λ，则可建立起拉格朗日函数与原问题（5.5）之间的联系，即：

$$\sup_{\lambda\geqslant 0} L(\beta;\lambda)=\begin{cases} f(\beta), & g_j(\beta)\leqslant 0,\quad j=1,\ldots,m\\ +\infty, & \text{其他}\end{cases} \tag{5.7}$$

则 $f^*=\inf_{\beta\in\mathbb{R}^p}\sup_{\lambda\geqslant 0}L(\beta;\lambda)$。习题 5.2 会详细描述这种等价性。

① 不等式（i）是可微函数 f 的凸性的等价定义；在 $\bar{\beta}\in C$ 处的一阶泰勒近似给出了切线到 f 的下界。

对于凸规划，引入拉格朗日函数可使求解约束问题（5.5）简化为求解等价的无约束问题。具体而言，对 f 和 $\{g_j\}$ 做出一些限制后，拉格朗日对偶理论可保证存在一个最优化的拉格朗日乘子向量 $\lambda^* \geqslant 0$，使 $f^* = \min_{\beta\in\mathbb{R}^p} L(\beta;\lambda^*)$。因此，问题（5.5）的最优解 β^* 除了满足可行约束条件 $g_i(\beta^*) \leqslant 0$，还必须让拉格朗日函数的梯度为零，即

$$0 = \nabla_\beta L(\beta^*;\lambda^*) = \nabla f(\beta^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(\beta^*) \tag{5.8}$$

当只有一个约束函数 g 时，这个条件就简化为 $\nabla f(\beta^*) = -\lambda^* \nabla g(\beta^*)$，它的直观几何解释如图 5-2 所示。实际上，在最优解 β^* 处，函数 f 的法向量 $\nabla f(\beta^*)$ 与约束曲线 $g(\beta)=0$ 的法向量方向相反。即在 β^* 处，函数 f 的法向量垂直于约束的切线。因此，若在 β^* 处沿着 $g(\beta)=0$ 切线的方向移动，不会减少 $f(\beta)$ 的值。

通常，Karush–Kuhn–Tucker（KKT）条件将最优的拉格朗日乘子向量（也称对偶向量）$\lambda^* \geqslant 0$ 与原问题的最优 $\beta^* \in \mathbb{R}^p$ 联系在一起。具体的 KKT 条件为

(a) **原问题可行**：$g_j(\beta^*) \leqslant 0$，$j = 1, \ldots, m$。

(b) **互补松弛**：$\lambda_j^* g_j(\beta^*) = 0$，$j = 1, \ldots, m$。

(c) **拉格朗日条件**：(β^*, λ^*) 满足条件式（5.8）。

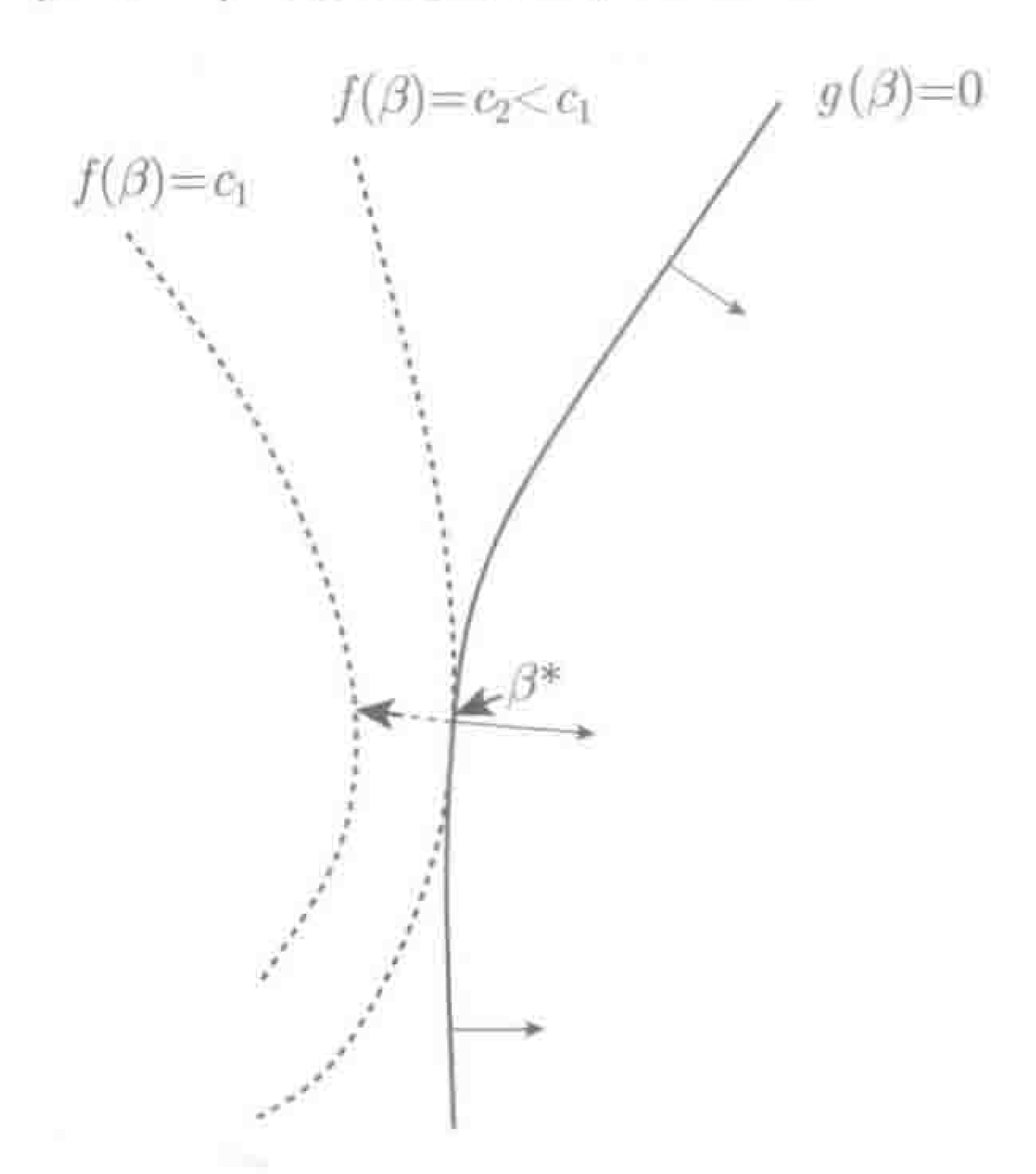


图 5-2 拉格朗日乘子方法示意。在只有一个约束条件 $g \leqslant 0$ 的情况下，最小化函数 f。在最优解 β^* 处，$\nabla f(\beta^*)$ 是目标函数 f 的水平集的法向量，与约束边界 $g(\beta)=0$ 在此处的法向量方向相反。因此，沿 $g(\beta)=0$ 移动不会减少 $f(\beta^*)$ 的值

只要优化问题满足**强对偶**条件，KKT 条件就是 β^* 为全局最优解的充分必要条件（详见习题 5.4）。由互补松弛条件可知：在得到最优解时，若约束 $g_j(\beta) \leqslant 0$ 不起作用，即 $g_j(\beta^*) < 0$，则乘子 λ_j^* 一定为零。因此，在互补松弛条件下，由拉格

朗日梯度条件可知：法向量 $-\nabla f(\beta^*)$ 是梯度向量 $\{\nabla g_j(\beta^*)|\lambda_j^* > 0\}$ 的线性组合，其组合系数为正。

5.2.2 非可微函数和次梯度

在一些实际的优化问题中，有的目标函数是凸函数但不可微。比如，ℓ_1 范数 $g(\beta) = \sum_{j=1}^{p} |\beta_j|$ 是一个凸函数，但它在某些点不可微，对于这些不可微的点，至少有某个坐标 β_j 等于零。对于这类问题，前面介绍的优化条件，即一阶条件式 (5.3) 和拉格朗日条件式 (5.8)，不能直接使用，因为这些条件会涉及目标函数和约束函数的梯度。但可以对凸函数的梯度概念进行推广，从而得到更一般的最优理论。

可微凸函数的基本性质是，一阶切线近似总是下界。次梯度的概念基于此观点的自然推广。对于一个凸函数 $f : \mathbb{R}^p \to \mathbb{R}$，如果不等式

$$f(\beta') \geqslant f(\beta) + \langle z, \beta' - \beta \rangle,\ \beta' \in \mathbb{R}^p \tag{5.9}$$

成立，则称向量 $\boldsymbol{z} \in \mathbb{R}^p$ 为 f 在 β 处的次梯度。就几何意义而言，次梯度向量 z 是支撑 f 的上方图（epigraph）的（非垂直）超平面的法向量。f 在 β 处的所有次梯度集合称为**次微分**，用 $\partial f(\beta)$ 表示。若 f 在 β 处可微，则次微分为一个向量，即 $\partial f(\beta) = \{\nabla f(\beta)\}$。对于不可微的点，次微分是一个包含所有次梯度的凸集。比如绝对值函数 $f(\beta) = |\beta|$ 的次微分为

$$\partial f(\beta) = \begin{cases} \{+1\}, & \beta > 0 \\ \{-1\}, & \beta < 0 \\ [-1, +1], & \beta = 0 \end{cases} \tag{5.10}$$

通常记 $z \in \operatorname{sgn}(\beta)$，即 z 属于绝对值函数在 β 处的次微分。

图 5-3 为函数 $f : \mathbb{R} \to \mathbb{R}$ 和两个点 β_1、β_2 处的次梯度的示意图。函数在 β_1 处可微，因此只有一个次梯度 $f'(\beta_1)$。函数在 β_2 处不可微，因此有多个次梯度，每个次梯度对应一个切平面，该切平面是 f 的下界。

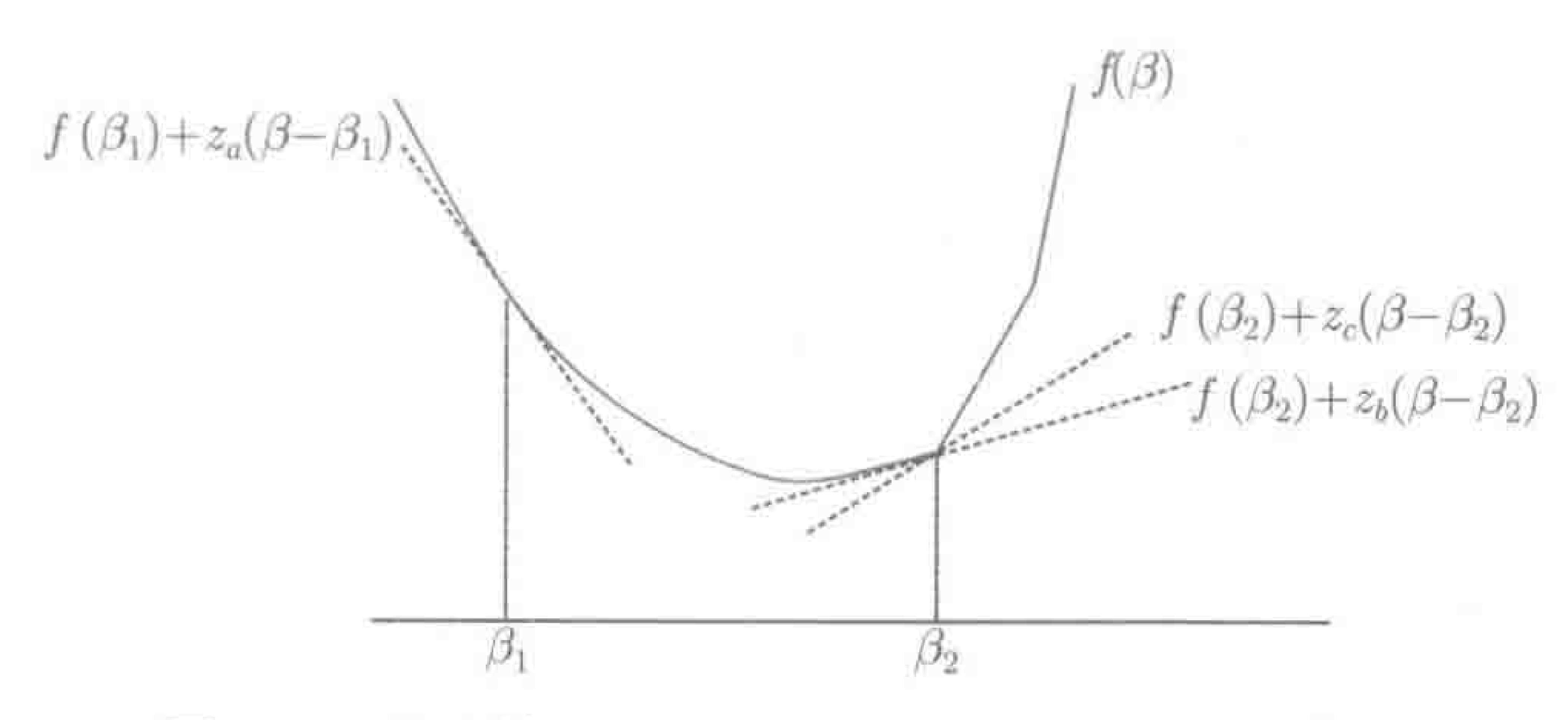


图 5-3　凸函数 $f : \mathbb{R} \to \mathbb{R}$，及 β_1、β_2 处的次梯度

次梯度有什么用呢？对于凸优化问题（5.5），假设有一个或多个不可微的凸函数 $\{f, g_i\}$，这时拉格朗日为零的条件式（5.8）不再有用。不过，也可对此函数放宽条件，得到广义的 KKT 理论

$$0 \in \partial f(\beta^*) + \sum_{j=1}^{m} \lambda_j^* \partial_{gj}(\beta^*) \tag{5.11}$$

上面的式子用次梯度替换了式（5.8）中 KKT 条件的梯度。次梯度是一个集合，因此式（5.11）说明零向量属于次梯度之和 ①。

例 5.1: Lasso与次梯度　求解可微凸目标函数 f 的最小值，约束条件仅为 $g(\beta) = \sum_{j=1}^{p} |\beta_j| - R$，其中 R 为正数，整个优化问题与式（5.5）一样。因此，约束 $g(\beta) \leqslant 0$ 等价于 $\boldsymbol{\beta}$ 在半径为 R 的 ℓ_1 球面或球内。回想前面讨论过绝对值函数的次梯度（5.10），式（5.11）变为

$$\nabla f(\beta^*) + \lambda^* z^* = 0 \tag{5.12}$$

其中，次梯度向量满足 $z_j^* \in \operatorname{sgn}(\beta_j^*)(j = 1, \ldots, p)$。当目标函数 f 为平方误差函数：$f(\beta) = \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\beta\|_2^2$，这个条件与式（2.6）等价。

例 5.2: 原子范数与次梯度　**原子范数**（nuclear norm）是矩阵空间的一个凸函数。给定矩阵 $\boldsymbol{\Theta} \in \mathbb{R}^{m \times n}$（其中 $m \leqslant n$），这样的矩阵总是可分解为 $\boldsymbol{\Theta} = \sum_{j=1}^{m} \sigma_j u_j v_j^{\mathrm{T}}$，其中 $\{u_j\}_{j=1}^{m}$ 和 $\{v_j\}_{j=1}^{m}$ 分别为左奇异向量和右奇异向量，这些向量分别属于 $\mathbb{R}^m$ 和 $\mathbb{R}^n$ 空间，且各自正交；非负数 $\sigma_j \geqslant 0$ 称为奇异值。这就是矩阵 $\boldsymbol{\Theta}$ 的**奇异值分解**（Singular-Value Decomposition，SVD）。**原子范数**等于所有奇异值相加之和，即 $\|\boldsymbol{\Theta}\|_* = \sum_{j=1}^{m} \sigma_j(\boldsymbol{\Theta})$。注意，这是 ℓ_1 范数的推广形式，因为对于任意的对角矩阵（方阵），原子范数会退化成对角线元素的 ℓ_1 范数。在第 7 章会看到原子范数在各种矩阵的近似和分解中很有用。矩阵 $\boldsymbol{\Theta}$ 的原子范数的次微分 $\partial \|\boldsymbol{\Theta}\|_*$ 由所有矩阵 $\boldsymbol{Z} = \sum_{j=1}^{m} z_j u_j v_j^{\mathrm{T}}$ 组成，其中每个标量 $z_j \in \operatorname{sgn}(\sigma_j(\boldsymbol{\Theta}))$，$j = 1, \ldots, m$。这里留一个习题给读者：用定义（5.9）验证这个结论。

5.3　梯 度 下 降

上面针对不同类型的凸规划问题介绍了相应的优化条件，下面介绍各种迭代算法来求解这些优化问题。本节重点介绍一阶优化算法，也就是说，这些方法只需要梯度（或次梯度）信息，不需要高阶梯度信息。在现代统计中，大规模问题采用一阶梯度方法能顺利求解。

5.3.1　无约束的梯度下降

我们从最简单的情况开始介绍：求可微凸函数 $f : \mathbb{R}^p \to \mathbb{R}$ 无约束条件下的最

① 这里定义两个 p 维实数的子集 A 和 B 相加：$A + B := \{\alpha + \beta | \alpha \in A, \beta \in B\}$

小值。在这种情况下，假设已经得到了全局最优解 $\beta^* \in \mathbb{R}^p$，而得到这个解的充分必要条件是梯度条件 $\nabla f(\beta^*) = 0$。梯度下降是一种求解不动点方程的迭代算法：它通过迭代等式

$$\beta^{t+1} = \beta^t - s^t \nabla f(\beta^t),\quad t = 0, 1, 2, \ldots \tag{5.13}$$

来得到迭代序列 $\{\beta^t\}_{t=0}^{\infty}$，其中 $s^t > 0$ 是步长。这种迭代的几何解释为：计算梯度可得到下降最大的方向 $-\nabla f(\beta^t)$，然后朝这个方向移动的步长为 s^t。

这类梯度下降方法一般都选择一个方向 $\Delta^t \in \mathbb{R}^p$，使得 $\langle \nabla f(\beta^t), \Delta^t \rangle < 0$，然后执行的迭代操作

$$\beta^{t+1} = \beta^t + s^t \Delta^t,\quad t = 0, 1, 2, \ldots \tag{5.14}$$

内积 $\langle \nabla f(\beta^t), \Delta^t \rangle < 0$ 的几何意义是：选择的方向 Δ^t 与最大下降方向的夹角小于 $90°$。梯度下降迭代式（5.13）是一种 $\Delta^t = -\nabla f(\beta^t)$ 的特例。对于 Δ^t，还有一些有趣的选择，比如**基于对角缩放的梯度下降**（diagonally-scaled gradient descent），即给定一个对角矩阵 $\boldsymbol{D}^t \succ \boldsymbol{0}$，所使用的梯度下降方向为 $\Delta^t = -(\boldsymbol{D}^t)^{-1} \nabla f(\beta^t)$。当函数在某些坐标方向上变化比其他坐标方向快得多时，这类对角缩放就很实用。更常见的是牛顿方法，这种方法要求目标函数要能二次连续可微，其下降方向为

$$\Delta^t = -(\nabla^2 f(\beta^t))^{-1} \nabla f(\beta^t) \tag{5.15}$$

其中，$\nabla^2 f(\beta^t)$ 是 f 的海森矩阵，并假设该矩阵可逆。牛顿方法是二阶方法，因为需要求一阶和二阶导数。实际上，牛顿方法的步长设为 1，会得到函数 f 在 β^t 处的最小二阶泰勒近似。在某些条件下，牛顿方法的二次收敛率极佳。但计算牛顿方法的下降方向（5.15）比一阶方法的开销要大。

迭代算法 [包括梯度下降迭代式（5.13）] 的另一个重要问题是如何确定步长 s^t。对于某些有具体结构的问题，可证明采用固定步长（即 $s^t = s$，$t = 0, 1, \ldots$），迭代也可以收敛，见习题 5.1。通常，仅仅选择步长让 $f(\beta^{t+1}) < f(\beta^t)$ 是不够的，随便选择步长可能会导致算法收敛非稳定点。幸运的是，现在有很多相对简单的步长选择规则可以保证迭代的收敛。

- **限制最小规则**：选择步长 $s^t = \underset{s \in [0,1]}{\arg\min} f(\beta^t + s\Delta^t)$。这种选择很直观，但是每次迭代都要解一个一维优化问题。
- **Armijo 规则**：也称回溯线搜索（backtracking line search）。给定参数 $\alpha \in (0, 0.5)$ 和 $\gamma \in (0, 1)$，并设初始化步长为 $s = 1$，重复 $s \leftarrow \gamma s$，直到满足下降条件

$$f(\beta^t + s\Delta^t) \leqslant f(\beta^t) + \alpha s \left\langle \nabla f(\beta^t), \Delta^t \right\rangle \tag{5.16}$$

通常 $\alpha = 0.5$，$\gamma = 0.8$ 比较合理。条件式（5.16）可解释为：这里将接受 $f(\beta)$ 减少为 α 的一小部分，这是由线性外推法预测得到的（如图 5-4 所示）。

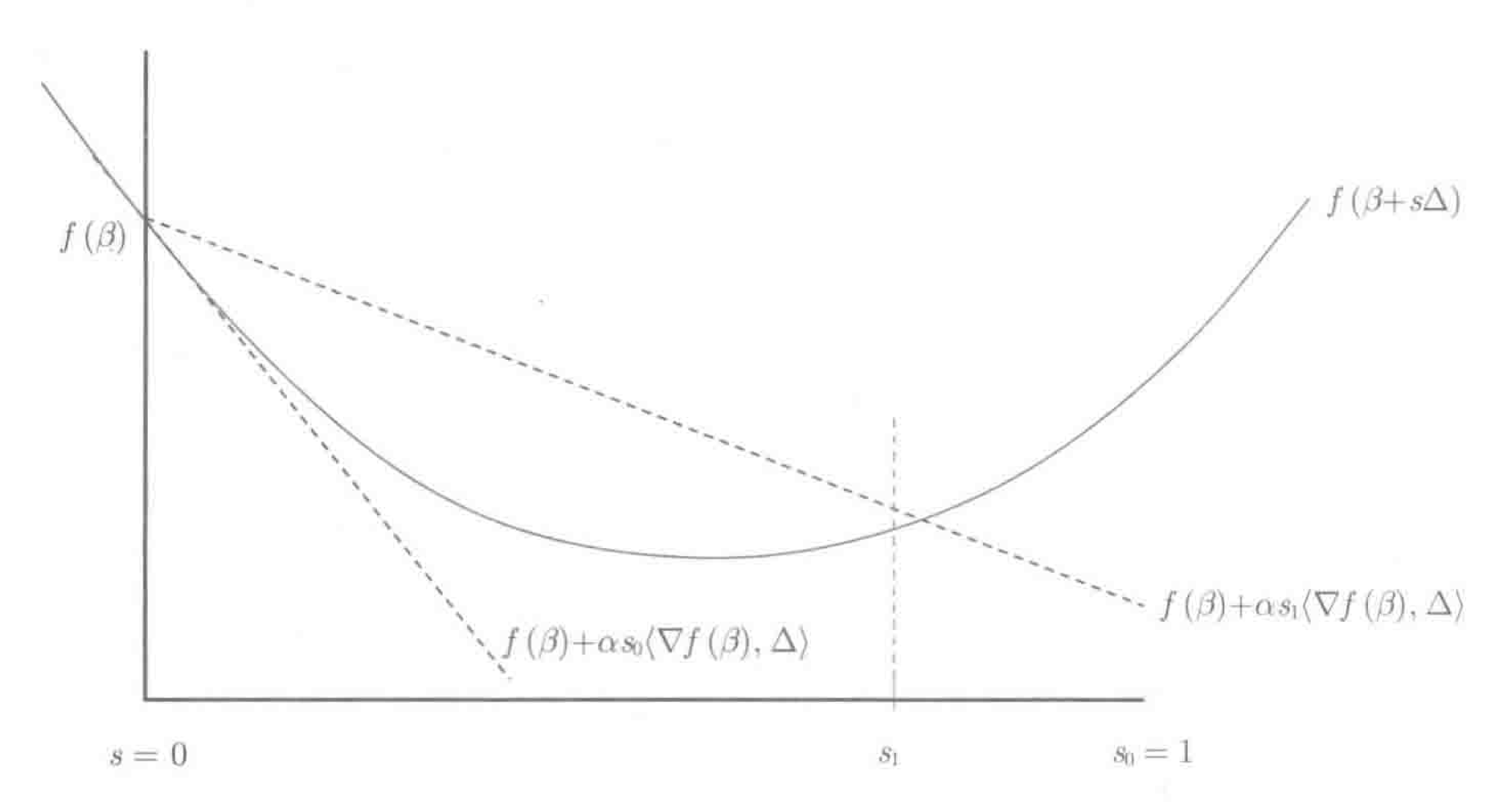


图 5-4　Armijo 规则，也称回溯线搜索。开始搜索时的步长 $s_0 = 1$，用一个因子 γ 来不断减少 s，直到满足 $f(\beta + s\Delta) \leqslant f(\beta) + \alpha s \langle \nabla f(\beta), \Delta \rangle$，即可得到 s_1

对于凸函数，这些步长选择方法与合适下降方向 $\{\Delta^t\}_{t=0}^{\infty}$ 相结合可收敛于凸函数 f 的全局最优解。进一步的讨论可参考文献注释。

5.3.2　投影梯度法

下面介绍带约束问题的梯度下降方法。为了让这些方法在几何上更加直观，有必要先看一下式（5.13）的另一种形式

$$\beta^{t+1} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ f(\beta^t) + \left\langle \nabla f(\beta^t), \beta - \beta^t \right\rangle + \frac{1}{2s^t} \|\beta - \beta^t\|_2^2 \right\} \tag{5.17}$$

这种形式可以看作在当前迭代中，将 f 线性化与平滑惩罚相加，并将结果最小化，平滑惩罚采用的是欧氏距离。

这种梯度下降的思想（针对无约束最小化的算法自然会得到投影梯度下降法）适用于有约束 $\beta \in \mathcal{C}$ 的最小化问题

$$\beta^{t+1} = \underset{\beta \in \mathcal{C}}{\arg\min} \left\{ f(\beta^t) + \left\langle \nabla f(\beta^t), \beta - \beta^t \right\rangle + \frac{1}{2s^t} \|\beta - \beta^t\|_2^2 \right\} \tag{5.18}$$

图 5-5 为该式子的示意图。这种方法先执行一个梯度相关步骤 $\beta^t - s\nabla f(\beta^t)$，然后将结果投影到凸约束集 $\mathcal{C}$ 上。只要计算投影相对简单，这其实是一种高效的算法。比如，给定一个 ℓ_1 球约束 $\mathcal{C} = \{\beta \in \mathbb{R}^p | \parallel \beta \parallel_1 \leqslant R\}$，通过软阈值法变换就能轻松计算这种投影。后面会详细介绍这方面的内容。

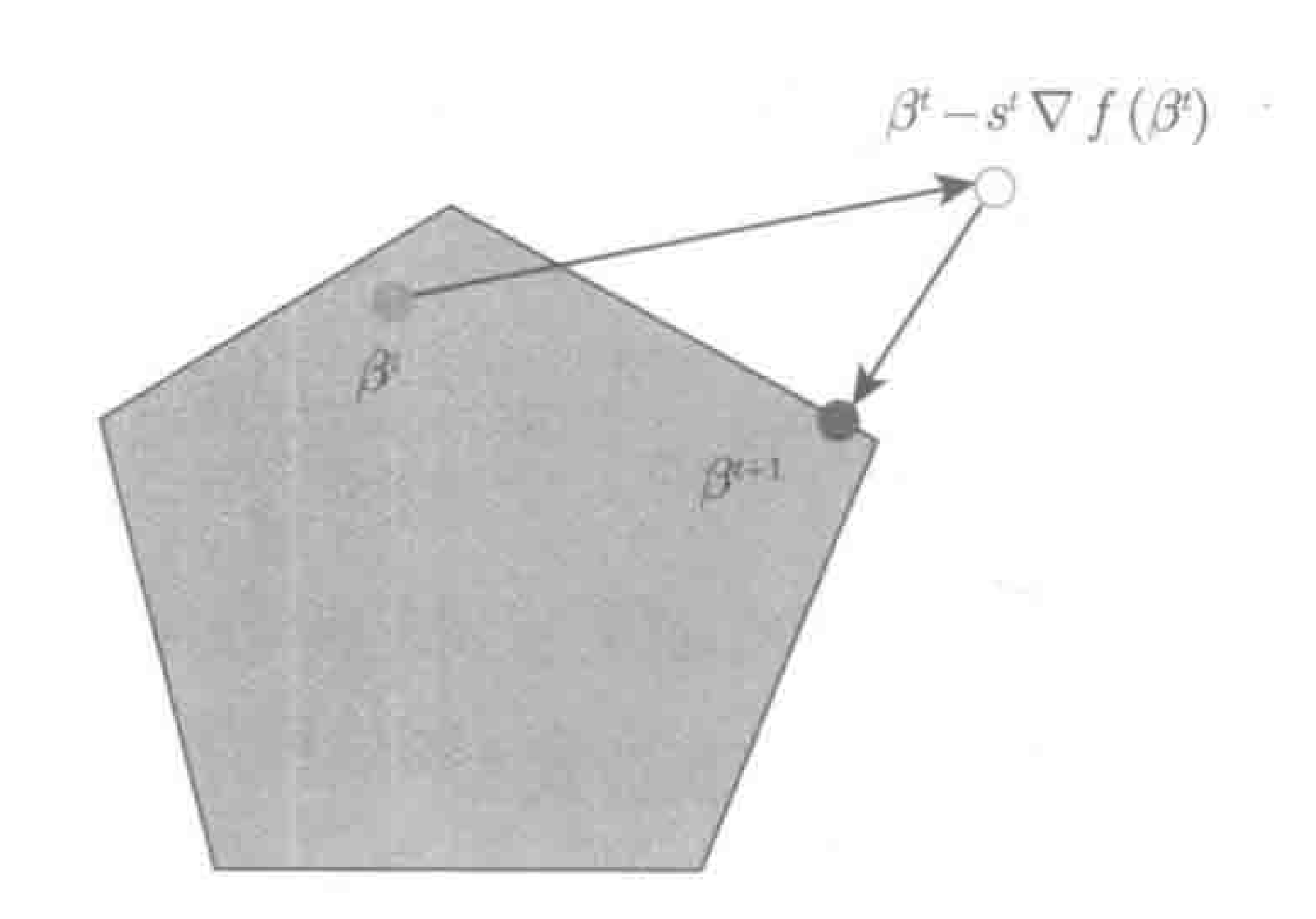


图 5-5　投影梯度的几何解释。当前迭代从 β^t 开始，按负梯度方向移动到 $\beta^t - s^t\nabla f(\beta^t)$，然后将结果通过欧几里得投影到凸约束集 $\mathcal{C}$ 上，得到下一次迭代的 β^{t+1}

5.3.3 近点梯度法

下面讨论投影梯度下降法的一般形式。正如之前所讨论的，很多目标函数 f 可以分解成两个函数之和：$f = g + h$，其中 g 是可微凸函数，h 为不可微凸函数。假设采用梯度类算法来求解目标函数的最小值，如何处理 f 中那部分不可微的函数 h 呢？

下面来看一下如何巧妙地解决这个难题。如前所述，通常的梯度下降可看作是将 f 的局部线性近似加上一个二次光滑项后，再对这种组合形式进行最小化 [见式（5.17）]。由这种思想可得到以下策略：通过线性化可微函数 g 来得到 f 的局部近似，并固定不可微函数。这会得到**广义的梯度迭代公式**，即

$$\beta^{t+1} = \mathop{\arg\min}_{\beta\in\mathbb{R}^p} \left\{ g(\beta^t) + \left\langle \nabla g(\beta^t), \beta - \beta^t \right\rangle + \frac{1}{2s^t}\|\beta - \beta^t\|_2^2 + h(\beta) \right\} \tag{5.19}$$

这个式子近似了可微部分 g，但不可微的 h 不变。

式（5.19）与投影梯度下降式（5.18）联系紧密。事实上，这可看成与拉格朗日函数类似的方式。为了更清楚地说明这种联系，可定义凸函数 h 的**近点映射**（proximal map），这是一种广义投影算子

$$\mathbf{prox}_h(\boldsymbol{z}) = \mathop{\arg\min}_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{z} - \theta\|_2^2 + h(\theta) \right\} \tag{5.20}$$

从这个定义可得出如下关系：

(a) $\mathbf{prox}_{sh}(z) = \mathop{\arg\min}_{\theta\in\mathbb{R}^p}\{\frac{1}{2s}\|z - \theta\|_2^2 + h(\theta)\}$

(b) 若

$$h(\theta) = I_{\mathcal{C}}(\theta) = \begin{cases} 0, & \theta \in \mathcal{C} \\ +\infty, & 其他 \end{cases}$$

则有 $\mathbf{prox}_h(z) = \arg\min_{\theta \in \mathcal{C}} \|z-\theta\|_2^2$，这就是 z 在集合 $\mathcal{C}$ 上的欧几里得投影。

(c) 若 $h(\theta) = \lambda \parallel \theta \parallel_1$，则 $\mathbf{prox}_h(z) = \mathcal{S}_\lambda(z)$，这是 z 的逐元素软阈值，见下面的例 5.3。

习题 5.7 会谈到，迭代式（5.19）与

$$\beta^{t+1} = \mathbf{prox}_{s^t h}(\beta^t - s^t \nabla g(\beta^t)) \tag{5.21}$$

等价。同理，很容易得到投影梯度的更新公式

$$\beta^{t+1} = \mathbf{prox}_{I_C}(\beta^t - s^t \nabla g(\beta^t)) \tag{5.22}$$

这个式子就是投影梯度步骤（5.18）。

只要近点映射的计算相对容易，更新式（5.21）的计算效率就会很高。对于统计中的一些问题，比如 ℓ_1 范数、组 lasso ℓ_2 范数、原子范数等，计算近点映射非常容易。相对于约束形式 $h(\theta) \leqslant R$，通常迭代式（5.21）更适合有正则化（有惩罚项）约束形式的统计问题。

例 5.3：基于 $\boldsymbol{\ell_1}$ 惩罚项的近点梯度下降　假设不可微的部分是（缩放）ℓ_1 惩罚项，即 $h(\theta) = \lambda \parallel \theta \parallel_1$。基于这种形式，在 t 次迭代、步长为 s^t 时，其近点梯度下降分成两个简单的步骤：

(1) 执行 $z = \beta^t - s^t \nabla g(\beta^t)$；

(2) 执行逐元素软阈值：$\beta^{t+1} = \mathcal{S}_{s^t\lambda}(\boldsymbol{z})$。

具体讲，近点映射式（5.21）为

$$\begin{aligned}\mathbf{prox}_{sh}(z) &= \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2s}\|z-\theta\|_2^2 + \lambda\|\theta\|_1 \right\} \\ &= \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2}\|z-\theta\|_2^2 + \lambda s\|\theta\|_1 \right\}\end{aligned} \tag{5.23}$$

这个优化问题有闭解形式，因为该目标函数可按坐标来分解成求和的形式

$$\frac{1}{2}\|z-\theta\|_2^2 + \lambda s\|\theta\|_1 = \sum_{j=1}^{p} \left\{ \frac{1}{2}(z_j - \theta_j)^2 + \lambda s|\theta_j| \right\} \tag{5.24}$$

从上式可看出，分别求解单变量问题就能解这个 p 维问题。这里留给读者一个习题：请验证，通过运用每一维如下形式的**软阈值算子** $\mathcal{S}_\tau : \mathbb{R}^p \to \mathbb{R}^p$，可以得到式（5.24）的解：

$$[\mathcal{S}_\tau(\boldsymbol{z})]_j = \text{sgn}(z_j)(|z_j| - \tau)_+ \tag{5.25}$$

其中，阈值 $\tau = s\lambda$。（这里的 $(x)_+$ 是 $\max\{x, 0\}$ 的简写。）

例 5.4：基于原子范数惩罚项的近点梯度下降　假设 h 为标量 λ 乘以原子范数。由前面的例 5.2 可知，原子范数将 $m \times n$ 的矩阵映射成实值函数，即 $\|\boldsymbol{\Theta}\|_* =$

$\sum_{j=1}^{m}\sigma_j(\boldsymbol{\Theta})$，其中 $\{\sigma_j(\boldsymbol{\Theta})\}$ 是矩阵 $\boldsymbol{\Theta}$ 的奇异值。在这种情况下，广义投影算子式（5.20）有形式

$$\mathbf{prox}_{sh}(\boldsymbol{Z})=\underset{\boldsymbol{\Theta}\in\mathbb{R}^{m\times n}}{\arg\min}\left\{\frac{1}{2s}\|\boldsymbol{Z}-\boldsymbol{\Theta}\|_F^2+\lambda\|\boldsymbol{\Theta}\|_\star\right\} \tag{5.26}$$

其中，$\|\boldsymbol{Z}-\boldsymbol{\Theta}\|_F^2$ 为 Frobenius 范数，定义为 $\|\boldsymbol{Z}-\boldsymbol{\Theta}\|_F^2=\sum_{j=1}^{m}\sum_{k=1}^{n}(Z_{jk}-\Theta_{jk})^2$，可看成是基于矩阵的欧几里得范数。虽然近点映射式 (5.26) 不再可分，但仍有相对简单的解（详见习题 5.8），可对 $\boldsymbol{Z}$ 进行奇异值分解，然后软阈值化这些奇异值，来得到迭代公式 $\Pi_{s,h}(\boldsymbol{Z})$。

对于目标函数 $f=g+h$，Nesterov（2007）给出了迭代式（5.21）收敛的充分条件。若函数 g 连续可微且具有 Lipschitz 梯度，即存在一个常数 L 使得

$$\|\nabla g(\beta)-\nabla g(\beta')\|_2\leqslant L\|\beta-\beta'\|_2,\quad \beta,\beta'\in\mathbb{R}^p \tag{5.27}$$

在这个条件下，若采用固定步长 $s^t=s\in(0,1/L]$，则存在一个常量 C 独立于迭代次数，于是式（5.21）满足

$$f(\beta^t)-f(\beta^*)\leqslant\frac{C}{t+1}\|\beta^t-\beta^*\|_2,\quad t=1,2\ldots \tag{5.28}$$

其中 β^* 为最优解。也就是说，第 t 次迭代得到的值 $f(\beta^t)$ 与最优值 $f(\beta^*)$ 之差会以收敛率 $\mathcal{O}(1/t)$ 减少。这个收敛率称为**次线性收敛率**（sublinear convergence），只要固定步长的取值在 $(0,1/L]$ 范围内，都会得到这样的收敛率。这个值选择的上界与 Lipschitz 常量 L 有关，可能存在，也可能不存在。实际上，Armijo 规则也会有跟式（5.28）一样的收敛率。图 5-6 给出了迭代次数与函数减少值之间的关系。

如果目标函数有其他结构，就可能得到更快的收敛率。比如，除了有式（5.27）的 Lipschitz 连续梯度，可微函数 g 还是**强凸**的，则存在一个 $\gamma>0$，使

$$g(\beta+\Delta)-g(\beta)-\langle\nabla g(\beta),\Delta\rangle\geqslant\gamma^2\|\Delta\|_2^2,\quad \beta,\Delta\in\mathbb{R}^p \tag{5.29}$$

成立。这个条件说明函数 g 在所有方向上至少拥有与二次函数 $\beta\to\gamma^2\|\beta\|_2^2$ 一样的曲率。结合条件（5.27）和（5.29），对于固定步长 $s\in(0,1/L]$，迭代式（5.21）能得到**线性或者几何**收敛率，即存在一个为正数常量 C 和收缩因子（contraction factor）$\kappa\in(0,1)$，使下式成立：

$$f(\beta^t)-f(\beta^*)\leqslant C\kappa^t\|\beta^0-\beta^*\|_2,\quad t=1,2\ldots \tag{5.30}$$

因此，在增加一个强凸条件后，$f(\beta^t)-f(\beta^*)$ 的差值会以某个 $\kappa\in(0,1)$ 按几何收敛率收敛。图 5-6 比较了线性收敛率和次线性收敛率 [见式（5.28）] 之间的不同。

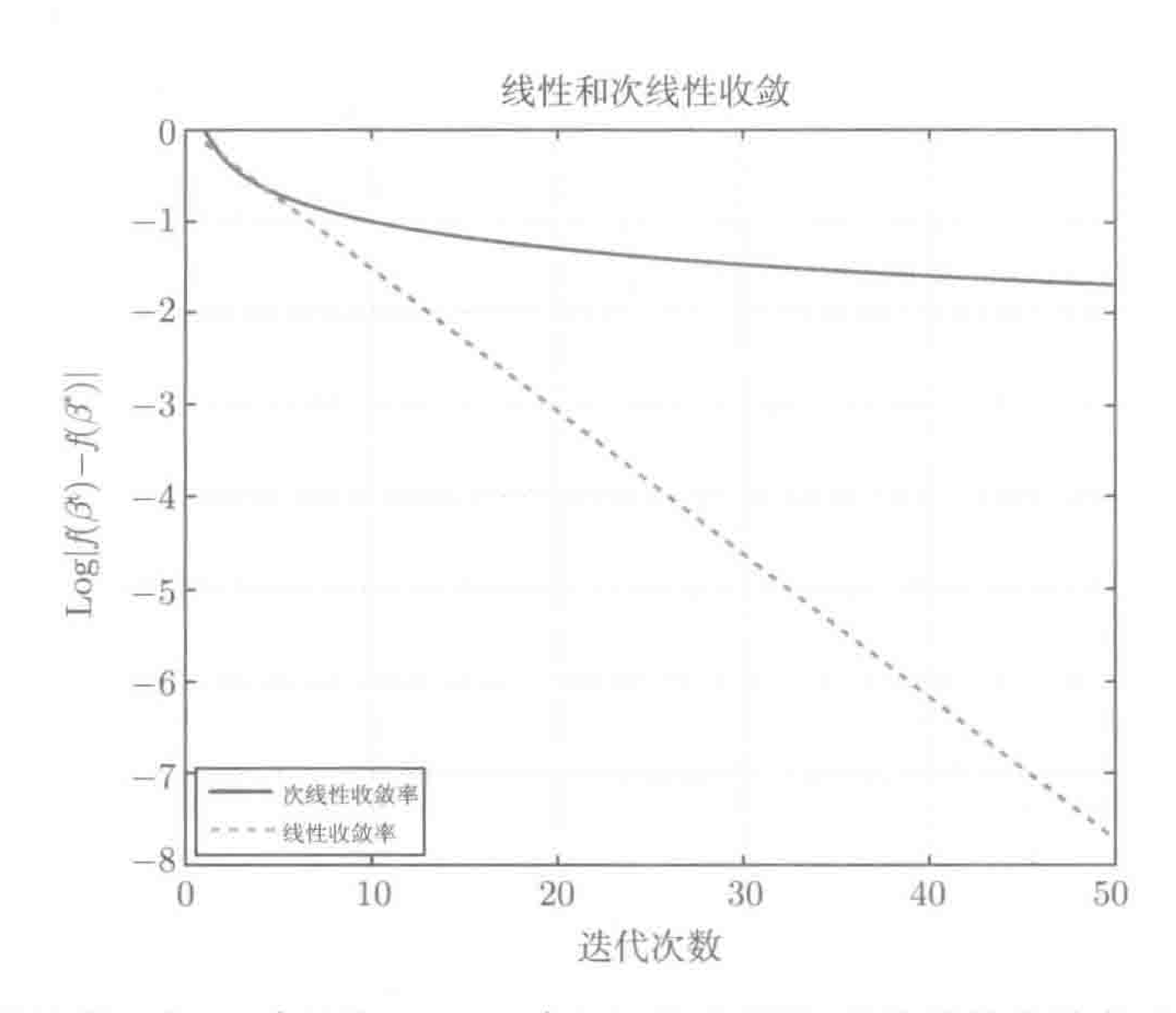


图 5-6　迭代次数 t 与 $\log\left|f(\beta^t)-f(\beta^*)\right|$ 之间的关系图。将次线性收敛率 [见式（5.28）] 与线性收敛率 [也称几何收敛率，见式（5.30）] 进行比较。对于有几何收敛率的算法，在对数尺度下的误差衰减是一条斜率为负的直线

例 5.5：基于 lasso 的近点梯度　对于 lasso 有

$$g(\beta)=\frac{1}{2N}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2,\quad h(\beta)=\lambda\|\beta\|_1$$

则近点梯度迭代式（5.21）的具体形式为

$$\beta^{t+1}=\mathcal{S}_{s^t\lambda}\left(\beta^t-s^t\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\beta^t)\right)\tag{5.31}$$

注意，这个更新式与基于坐标下降的更新式（见 5.4 节）很类似，尤其是步长 $s=1$，且假定输入数据经过归一化处理。这两个迭代过程形式相同，一个是单坐标上循环方式，而另一个（近似点梯度）是在所有坐标上同时进行。目前并不清楚哪种方式更有效。坐标下降的迭代过程可以利用系数向量的稀疏性，而且不用考虑优化步长；而近点梯度由于同时操作所有参数，因此可能更有效。它可能在一些问题中具有速度优势，因为一个向量同时乘以 $\boldsymbol{X}$ 和 $\boldsymbol{X}^{\mathrm{T}}$ 可以快速完成，例如快速傅里叶变换。Lipschitz 常量 L 是 $\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{N}$ 的最大特征值。可使用一个 $(0,1/L]$ 范围内的固定步长或采用回溯步选择形式。5.5 节会从数值方面来比较这两种方式。

5.3.4　加速梯度方法

这节会介绍由 Nesterov（2007）提出的加速梯度方法。假设 f 为可微凸函数，若用式（5.13）这样的标准梯度迭代公式，有可能在迭代路径呈现“锯齿形”，这是

一种不妙的情况，可能会减缓收敛率。为了解决这个问题，Nesterov（2007）提出了加速梯度方法，这种方法会将上一步的梯度方向和当前的梯度方向通过加权方式结合起来。

具体而言，加速梯度方法会使用一对序列 $\{\beta^t\}_{t=0}^{\infty}$ 和 $\{\theta^t\}_{t=0}^{\infty}$，以及一些初始值 $\beta^0=\theta^0$。这对序列可按

$$\beta^{t+1}=\theta^t-s^t\nabla f(\theta^t) \tag{5.32a}$$

$$\theta^{t+1}=\beta^{t+1}-\frac{t}{t+3}(\beta^{t+1}-\beta^t) \tag{5.32b}$$

进行迭代，其中 $t=0,1,2,\ldots$。若非光滑函数 f 能分解成"光滑函数 + 非光滑函数"的形式（即 $g+h$），Nesterov 的加速方法可与近点梯度迭代相结合，即采用下式代替式（5.23a）：

$$\beta^{t+1}=\mathbf{prox}_{s^t h}(\theta^t-s^t\nabla g(\theta^t)) \tag{5.33}$$

对于任何一种情形，其步长 s^t 要么是某个固定值，要么采用某种回溯线性搜索进行选择。

例 5.6: 具有动能 (momentum) 的近点梯度下降 考虑用近点梯度与加速方法相结合来求解基于 ℓ_1 正则化的 lasso 问题。根据式（5.31），可得到加速迭代方式

$$\begin{aligned}\beta^{t+1}&=\mathcal{S}_{s^t\lambda}\left(\theta^t+s^t\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\theta^t)\right)\\ \theta^{t+1}&=\beta^{t+1}+\frac{t}{t+3}(\beta^{t+1}-\beta^t)\end{aligned} \tag{5.34a}$$

这种迭代方式与 Beck and Teboulle（2009）提出的快速迭代软阈值算法（Fast Iterative Soft-thresholding Algorithm，FISTA）是等价，只是加速权重有一些细微差别。

为了研究这种迭代的原理，这里生成一个具有 $N=1000$ 个样本，每个样本有 $p=500$ 个特征的数据集。特征 x_{ij} 服从标准高斯分布，它们之间的两两相关系数为 0.5。系数 β_j 是一个 500 维的向量，其中 20 个元素不为零，非零位置服从标准高斯分布，可选择 1 到 p 之间的任意一个位置。图 5-17 给出了在两个不同的正则化参数 λ 下，广义梯度与 Nesterov 动能法性能的比较结果。两个算法使用的固定步长都为 s^t（取 $\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 最大特征值的倒数）。这里也采用了回溯线性搜索近似，其中 $s^t\in[0,0.5]$。可以看出，Nesterov 动能法相比广义梯度法在效率上有明显的提高，回溯线性搜索比固定步长的收敛效率要高。后面这种比较甚至没有考虑计算 $\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的最大特征值的成本：回溯线性搜索在合适的时候会使用较大的步长来加速计算。这里通过迭次数而不是时间来计算性能，但 Nesterov 动能法迭代步骤时

间要比广义梯度法的稍长一点。同时需注意，Nesterov 动能法对应的相对误差与迭代次数之间并非严格的单调递减关系。

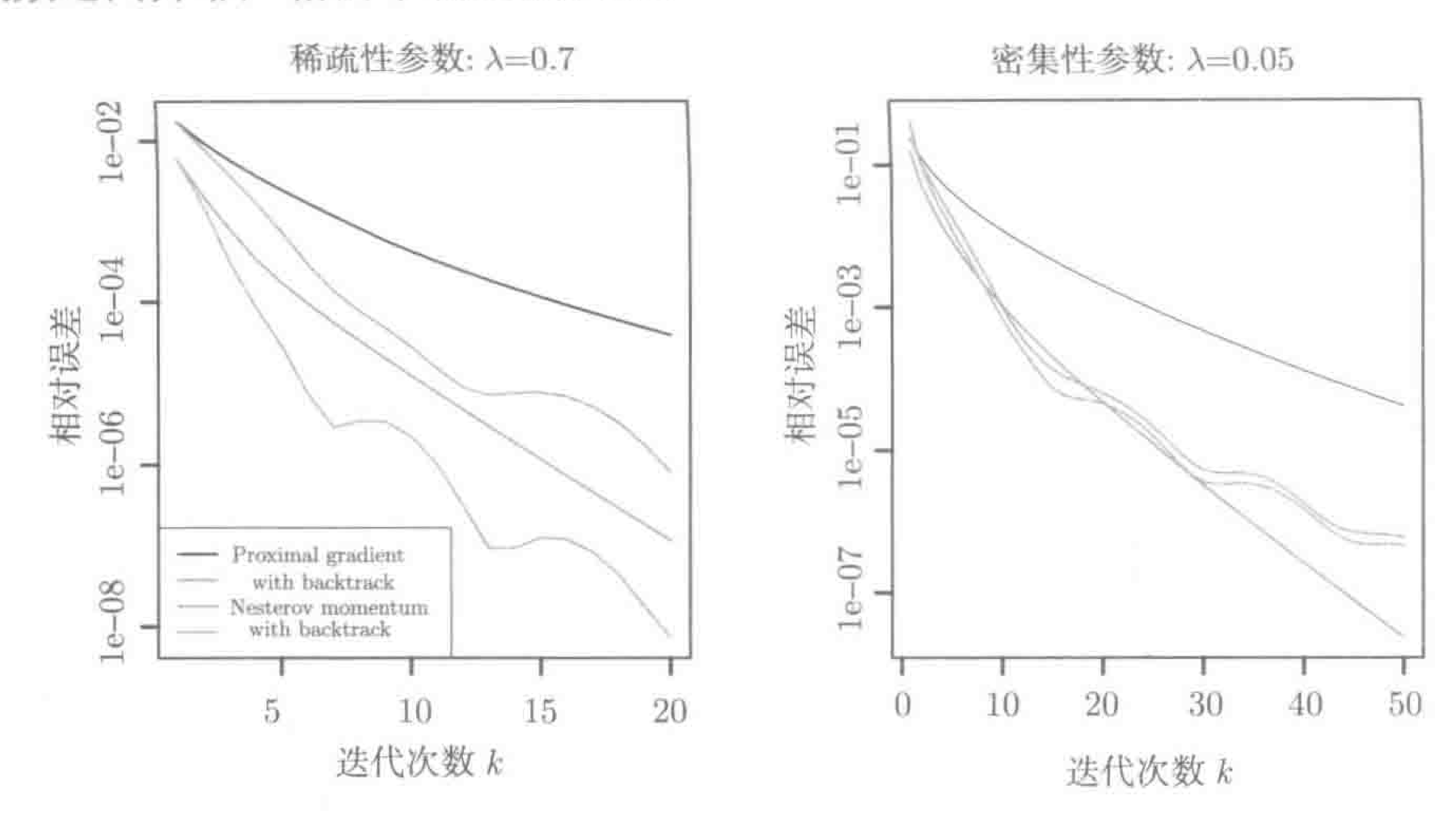


图 5-7　广义梯度及 Nesterov 动能法求解 lasso 问题时的性能比较。纵轴是通过误差度量 $\frac{[f(\beta^t)-f(\beta^*)]}{f(\beta^*)}$ 得到的相对误差，其中 β^* 为目标函数的最优解，β^t 是在迭代 t 次后的解。左图中的向量 β^* 含有 500 个系数，但只有 20 个不为 0；右图有 237 个系数不为零

式（5.32）和式（5.33）的计算量只比普通梯度迭代稍多。但 Nesterov（2007）证明这种变化会让收敛率大大提高：若设 g 满足 Lipschitz 条件式（5.27），则存在一个常量 $C>0$，使不等式

$$f(\beta^t)-f(\beta^*)\leqslant\frac{C}{(t+1)^2}\|\beta^0-\beta^*\|_2 \tag{5.35}$$

成立。因此，$f(\beta^t)-f(\beta^*)$ 的误差收敛率为 $\mathcal{O}(\frac{1}{t^2})$，而非加速方法的收敛率 $\mathcal{O}(\frac{1}{t})$ [见式（5.28）] 时。当 g 为强凸 [见式（5.29）] 时，虽然与之相关的收敛会涉及一个较小的收缩因子 κ，但它仍有不错的线性收敛率 [见式（5.30）]。更准确地讲，非加速方法收敛性跟一个收缩因子有关，该因子由 g 的条件数决定，而加速方法的收敛率由这个条件数的平方根决定。

5.4　坐标下降

某些类型的问题，包括 lasso 问题及其变形，具有可分性质，这自然就会引出基于坐标的最小化算法。**坐标下降**是一种迭代算法，在单个坐标方向执行 β^t 到 β^{t+1} 的迭代，然后在该坐标方向上求单变量最小值。更准确地讲，如果第 t 次迭代选择了坐标 k，则可以采用

$$\beta_k^{t+1} = \arg\min_{\beta_k} f(\beta_1^t, \beta_2^t, \ldots, \beta_{k-1}^t, \beta_k^t, \beta_{k+1}^t, \ldots, \beta_p^t) \tag{5.36}$$

进行迭代，其中 $\beta_j^{t+1} = \beta_j^t$，$j \neq k$。通常迭代过程中会按固定顺序选择坐标。这种方式也可推广到**块坐标下降法**（block coordinate descent）。同组 lasso 相似，在块坐标下降法中，变量被划分成不重叠的块，每次迭代会在单块上执行最小化。

5.4.1 可分性和坐标下降

坐标下降什么时候会收敛到凸函数的全局最小值？一个充分但有些严苛的条件是 f 连续可微，并在每个坐标方向上都严格为凸。但各种正则化所得到的优化问题并不需要可微。使用坐标下降来求解这类问题时会面临更多的问题。从下面的介绍可以看到，这类问题通常得不到最优化解。对于满足可分条件的这类最优化问题，用坐标下降可得最优化解，尤其是目标函数 f 有分解形式

$$f(\beta_1, \ldots \beta_p) = g(\beta_1, \ldots \beta_p) + \sum_{j=1}^{p} h_j(\beta_j) \tag{5.37}$$

其中，$g : \mathbb{R}^p \to \mathbb{R}$ 是可微凸函数，单变量函数 $h_j : \mathbb{R} \to \mathbb{R}$ 是凸函数（但不必是可微函数）。标准的 lasso 问题（2.5）就是具有这种结构的一个重要例子，其 $g(\beta) = \frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$，$h_j(\beta_j) = \lambda \cdot |\beta_j|$。Tseng（1988, 2001）证明：对于任何有可分结构式（5.37）的凸损失函数 f，坐标下降算法（5.36）都可得到全局最优解。这一结论的关键在于不可微函数 $h(\beta) = \sum_{j=1}^{p} h_j(\beta_j)$ 具有可分性，即 $h(\beta)$ 由单变量函数相加得到。这也说明坐标下降法适合 lasso 和本书所介绍的其他问题。同时也说明若 h 不可分，坐标下降法不会保证可以收敛，这种情况下有可能得不到全局最优解。

例 5.7: 不适合坐标下降法的例子 下面介绍一个不符合式（5.37）的例子，即 4.5 节介绍的融合 lasso，它是不可微函数 $h(\beta) = \sum_{j=1}^{p} |\beta_j - \beta_{j-1}|$。图 5-8 用来说明该问题的难度。该图所示为一个融合 lasso 问题，有 100 个参数，它的解为参数 $\beta_{63} = \beta_{64} \approx -1$。左图和中图分别给出函数 $f(\beta)$ 在除 β_{63} 和 β_{64} 以外，将其他参数都设置为全局最小值时的函数关系图。从图可以看到，对这两种情形分别采用坐标下降法，它们都会在各自的拐点处停下来，在单坐标方向上不动。必须同时移动 β_{63} 和 β_{64} 才能得到最小值。

Tseng（2001）给出了更一般、更直观的坐标下降法收敛条件。这个条件依赖于目标函数 f 的方向导数。给定一个方向 $\Delta \in \mathbb{R}^p$，在 β 处的方向导数下界为

$$f'(\beta; \Delta) := \liminf_{s \downarrow 0} \frac{f(\beta + s\Delta) - f(\beta)}{s} \tag{5.38}$$

粗略地讲，坐标下降方法仅能获取关于方向 $\boldsymbol{e}^j = (0, 0, \ldots, e_j, 0, \ldots, 0)$ 的信息，其中 $e_j \in \mathbb{R}$。因此，可认为坐标下降算法得到的点 β 满足条件

$$f'(\beta; \boldsymbol{e}^j) \geqslant 0 \text{ ，} \quad j = 1, \ldots, p \text{ ，} \quad \boldsymbol{e}^j \text{ 为坐标向量} \tag{5.39}$$

对于这样的点，再也没有进一步可减少函数值的坐标方向，因此，任意 β 都需要满足式（5.39），而且对于任意方向 $\Delta \in \mathbb{R}^p$，需满足 $f'(\beta, \Delta) \geqslant 0$。Tseng（2001）称这种条件为**正则性**（regularity）。这个条件就排除了图 5-8 的情形。在图 5-8 中，沿着所有坐标方向移动都有 $f'(\beta; \Delta) \geqslant 0$，但不沿坐标方向而是沿其他方向就不会有 $f'(\beta; \Delta) \geqslant 0$。除此以外还需注意，虽然这种可分但不可微的目标函数会有正则性，但不可微也不可分的函数也会有这种正则性。比如函数

$$h(\beta_1, \ldots, \beta_p) = |\beta|^{\mathrm{T}} \boldsymbol{P} |\beta| = \sum_{j,k=1}^{p} |\beta_j| P_{jk} |\beta_k| \tag{5.40}$$

其中 $\boldsymbol{P}$ 是对称正定矩阵。

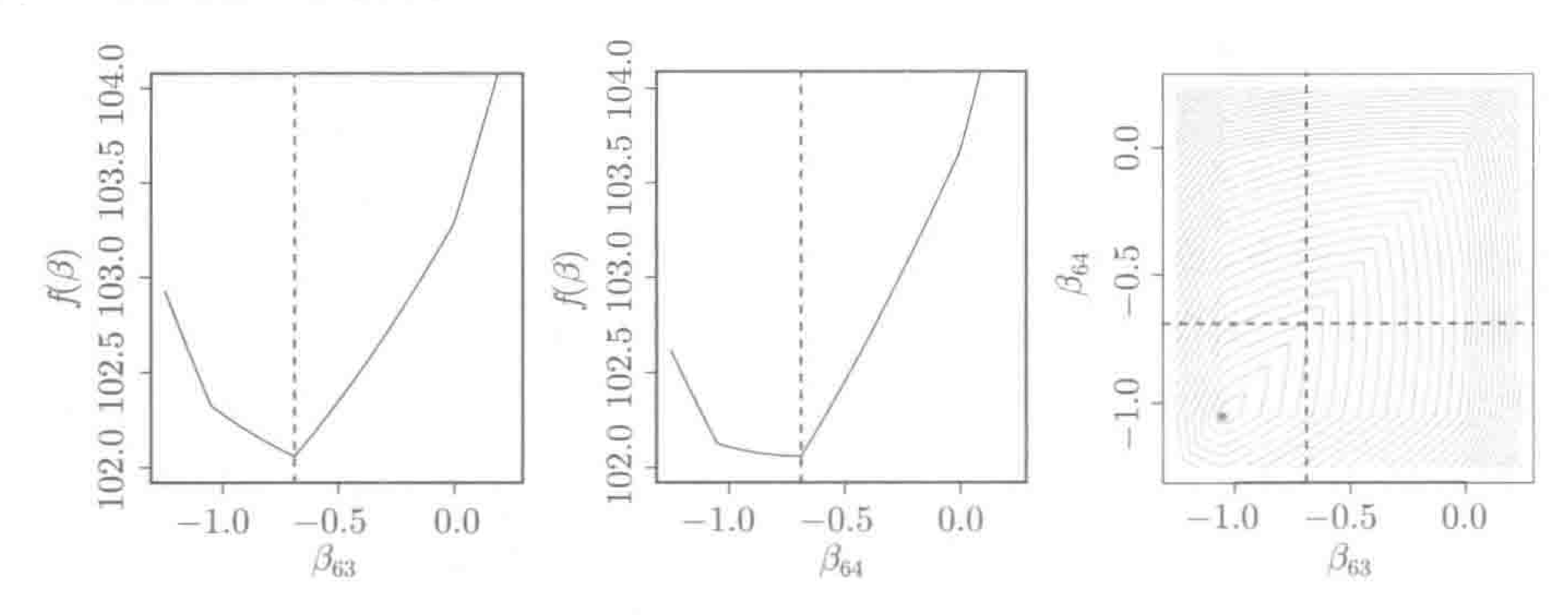


图 5-8　融合 lasso 问题不适合使用坐标下降法，图中给出的融合 lasso 有 100 个参数，但只针对两个参数 β_{63} 和 β_{64} 进行优化，当它们同时取 -1.05 时会得到目标函数 f 的最小值，即右图那个实心点。左图和中图分别是将 β_{63} 和 β_{64} 作为函数的变量，其他参数都设为全局最小值时的函数关系图。从这两幅图可以看出，在 β_{63} 和 β_{64} 分别都取 -0.69 而不是 -1.05 时，函数获得最小值。右图是二维平面的轮廓图。采用坐标下降在点 $(-0.69, -0.69)$ 处得到最小值。虽然这是一个严格凸的问题，但采用坐标下降法会卡在非最小值处。为了得到最小值，必须要同时移动 β_{63} 和 β_{64}

5.4.2　线性回归和 lasso

对于第 2 章的优化问题（2.5），得到最优解的条件是

$$-\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right) x_{ij} + \lambda s_j = 0 \tag{5.41}$$

其中 $s_j \in \operatorname{sgn}(\beta_j)$，$j = 1, 2, \ldots, p$。通过循环方式，可用坐标下降算法来简单地求解这些方程，其中 $j = 1, 2, \ldots, p, 1, 2, \ldots$。

由于截距 β_0 对求解没有影响，可分别通过响应向量 y_i 和协变量向量 x_i 的均值来对它们进行中心化，从而在计算 β_j 时忽略截距。(当然，在 OLS 中，截距最后可由 $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} x_j \hat{\beta}_j$ 计算得到。) 为了简化问题，这里定义**部分残差** $r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k$。从这个定义可看出，它会去掉除第 j 个输入变量的所有当前拟和结果。所以针对 $\boldsymbol{\beta}_j$ 的解满足

$$\hat{\beta}_j = \frac{\mathcal{S}_\lambda \left(\frac{1}{N} \sum_{i=1}^{N} r_i^{(j)} x_{ij} \right)}{\frac{1}{N} \sum_{i=1}^{N} x_{ij}^2} \tag{5.42}$$

其中 $\mathcal{S}_\lambda(\theta) = \operatorname{sgn}(\theta)(|\theta| - \lambda)_+$ 是软阈值算子。如果除了进行中心化还对变量进行归一化，使其样本方差为 1（这是种不错的做法，尤其是当变量的单位不一样时），则这种迭代会具有特别简洁的形式

$$\hat{\beta}_j = \mathcal{S}_\lambda \left(\tilde{\beta}_j \right) \tag{5.43}$$

其中 $\tilde{\beta}_j$ 是关于第 j 个变量的部分残差的线性回归系数。如果线性回归采用的是弹性网惩罚项 $(1-\alpha)\beta_j^2/2 + \alpha|\beta_j|$，则迭代式（5.42）会变成

$$\widehat{\beta}_j = \frac{\mathcal{S}_{a\lambda} \left(\frac{1}{N} \sum_{i=1}^{N} r_i^{(j)} x_{ij} \right)}{\frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 + (1-\alpha)\lambda} \tag{5.44}$$

或者标准形式下的

$$\widehat{\beta}_j = \frac{\mathcal{S}_{a\lambda}(\widetilde{\beta}_j)}{1 + (1-\alpha)\lambda} \tag{5.45}$$

有很多策略可让这些运算变得高效。为了让表示更简洁，假设输入数据进行了归一化，即使其样本均值为 0，方差为 1。对于没有归一化的数据，也有类似的步骤。

部分残差。注意：$r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k = r_i + x_{ij} \hat{\beta}_j$，其中 r_i 表示第 i 个样本当前的残差。由于对训练样本 $\{x_j\}_{j=1}^{p}$ 进行了归一化，于是有

$$\frac{1}{N} \sum_{i=1}^{N} x_{ij} r_i^{(j)} = \frac{1}{N} \sum_{i=1}^{N} x_{ij} r_i + \widehat{\beta}_j \tag{5.46}$$

从这种表示可看出坐标下降的计算效率很高。因为有许多系数为零，并且在阈值化后也是如此，所以没有什么需要改变。式（5.46）的主要计算成本为计算求和，其时间复杂度为 $\mathcal{O}(N)$。另外，如果阈值化后系数发生了变化，则 r_i 的时间复杂度变为

$\mathcal{O}(N)$，因此，这一步的时间复杂度为 $\mathcal{O}(2N)$。p 个变量完整循环的时间复杂度为 $\mathcal{O}(pN)$。由于这种方式能直接利用数据的内积，因此 Friedman et al.（2010b）称其为**简单更新**（naive updating）。

协方差更新 当 $N \gg p$ 且 N 很大时，**协方差更新**（covariance updating）通常要比简单更新的效率高一些。若乘以 $1/N$，则式（5.46）右边的第一项可写为

$$\sum_{i=1}^{N} x_{ij} r_i = \langle \boldsymbol{x}_j, \boldsymbol{y} \rangle - \sum_{k \| \widehat{\beta_k} | > 0} \langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle \widehat{\beta_k} \tag{5.47}$$

基于这个式子，可将每个特征与 $\boldsymbol{y}$ 做内积。对每次加到模型里的新特征 $\boldsymbol{x}_k$ 可计算并存储它与其他特征之间的内积，其时间复杂度为 $\mathcal{O}(Np)$。我们也会保存式（5.47）中的 p 个梯度分量。如果当前模型的某个系数发生变化，则更新每个梯度的时间复杂度为 $\mathcal{O}(p)$。因此，模型有 k 个非零项的情况下，如果没有新的变量变成非零，整个循环的时间复杂度为 $\mathcal{O}(pk)$；对于模型中新加入的每个变量，时间复杂度为 $\mathcal{O}(Np)$。有一点很重要，不是每一步都会做 $\mathcal{O}(N)$ 的计算。

热启动。通常需要一系列 lasso 的解，这意味着要有一个递减序列 $\{\lambda_\ell\}_0^L$。容易看出，所需要的最大值为

$$\lambda_0 = \frac{1}{N} \max_j |\langle \boldsymbol{x}_j, \boldsymbol{y} \rangle| \tag{5.48}$$

这是因为任何更大的值都会得到一个空模型。一种方式是按对数尺度来创建从 λ_0 到 $\lambda_L = \epsilon\lambda_0 \approx 0$ 的递减序列 $\{\lambda_\ell\}_0^L$，R 包 glmnet 采用的就是这种方式。对于解 $\hat{\beta}(\lambda_{\ell+1})$，解 $\hat{\beta}(\lambda_l)$ 通常是很好的热启动。而且从 0 开始时，其 $\ell = 0$，非零元素数目会缓慢增加。将 $L = 100$ 增加为 $2L$，其计算时间不会成倍增加，因为使用了热启动，每次只需要很少的迭代。

活动集收敛。在新值 λ_ℓ 处对一组有 p 个变量的集合进行单次迭代之后，从热启动解 $\hat{\beta}(\lambda_{\ell-1})$ 开始，可定义活动集 $\mathcal{A}$ 来保存当前有非零系数的变量，因此可让迭代算法仅使用 $\mathcal{A}$ 中的变量。收敛时可将所有忽略的变量检验一遍。如果所有变量都通过简单的检验 $\frac{1}{N}|\langle \boldsymbol{x}_j, \boldsymbol{r} \rangle| < \lambda_\ell$（其中 $\boldsymbol{r}$ 为当前的残差），就会得到整个集合（p 个变量）的解。没有通过检验的变量可放到 $\mathcal{A}$ 中，然后再重复这个过程。实际上，在所有变量放到 $\mathcal{A}$ 中之前，这个迭代过程会维护一个不断更新的活动集，其中包含在迭代过程中出现的任何有非零系数的变量。

强集收敛。可定义一个与上面活动集类似的变量子集。设 $\boldsymbol{r}$ 为 $\hat{\beta}(\lambda_{\ell-1})$ 处的残差，需要计算在 λ_ℓ 时的解。先定义一个强集 $\mathcal{S}$ 为

$$\mathcal{S} = \left\{ j \| \frac{1}{N} \langle \boldsymbol{x}_j, \boldsymbol{r} \rangle | > \lambda_\ell - (\lambda_{\ell-1} - \lambda_\ell) \right\} \tag{5.49}$$

现在仅用 $\mathcal{S}$ 中的变量来求解。在少数情况下，强集将包含最优活动集。强规则非常有用，特别是当 p 是非常大（10 万，甚至上百万）时。5.10 节会进行这方面的讨论。

稀疏性。在上面讨论的内容中，主要的运算是 N 维向量对之间的内积，其中至少一个是矩阵 $\boldsymbol{X}$ 中的列。如果 $\boldsymbol{X}$ 稀疏，就可以高效地计算这些内积。比如文档分类，它的特征向量通常为词袋（bag of word）模型。根据文档每个单词是否存在于字典来对该文档进行评分（有时会采用单词出现的次数或这些次数的一些变换来评分）。因为大多数单词都不存在，所以文档的特征向量元素大多为零，所以由这些向量构成的矩阵有大量元素为零。可按稀疏列有效存储该矩阵，即只存储非零元素及它们出现的位置。在计算内积时，仅把非零元素相乘之后加在一起即可。

惩罚强度。模型的公式在默认情况下会对每一项采用相同的惩罚参数 λ。第 j 个变量对应的惩罚强度为 $\gamma_i \geqslant 0$，所有的惩罚加在一起为

$$\lambda \sum_{j=1}^{p} \gamma_j P_\alpha(\beta_j) \tag{5.50}$$

有一些 γ_i 为零，这说明它们不会惩罚对应的变量。

参数的界。坐标下降法允许对每个参数设置上界和下界，即

$$L_j \leqslant \beta_j \leqslant U_j \tag{5.51}$$

一般情况下，$-\infty \leqslant L_j \leqslant 0 \leqslant U_j \leqslant \infty$。但有时需要约束所有系数为非负。在计算坐标更新后，如果参数超过所设定的界，就可将其设为最接近的边界。

5.4.3 逻辑斯蒂回归和广义线性模型

前面介绍了平方误差损失函数，下面介绍其他的指数族，即**广义线性模型**。简单起见，我们关注最重要的一类广义线性模型，即逻辑斯蒂回归。逻辑斯蒂回归的输出为二值类型，用 G 表示逻辑斯蒂回归的输出值（也称类标签），它的取值为 -1 或 1。标准的逻辑斯蒂回归用对数形式的线性模型来表示类概率：

$$\log \frac{\Pr(G=-1|x)}{\Pr(G=1|x)} = \beta_0 + x^{\mathrm{T}}\beta \tag{5.52}$$

详见 3.2 节该公式的介绍。

正则化的最大似然函数会拟和逻辑斯蒂回归模型。对第 i 个样本出现的概率用 $p\,(x_i;\beta_0,\beta) = \Pr(G=1|x_i)$ 来表示，最大化带惩罚项的对数似然函数为

$$\frac{1}{N}\sum_{i=1}^{N}\left\{I(g_i=1)\log p(x_i;\beta_0,\beta) + I(g_i=-1)\log(1-p(x_i;\beta_0,\beta))\right\} - \lambda P_\alpha(\beta) \tag{5.53}$$

其中 $y_i = I(g_i = -1)$，式（5.53）的对数似然部分可以重写成更简洁的形式

$$\ell(\beta_0, \beta) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot (\beta_0 + x_i^{\mathrm{T}}\beta) - \log(1 + \mathrm{e}^{\beta_0 + x_i^{\mathrm{T}}\beta})\right] \tag{5.54}$$

该函数为凹函数。用牛顿方法求解没有惩罚的对数似然函数（5.54），就相当于迭代最小二乘。如果当前的参数估计为 $(\tilde{\beta}_0, \tilde{\beta})$，则可对当前估计进行二阶泰勒展开。若令 $\tilde{p}(x_i) = p\left(x_i; \beta_0, \beta\right)$，$w_i = \tilde{p}(x_i)(1 - \tilde{p}(x_i))$，则泰勒展开会得到二次目标函数

$$\ell_Q(\beta_0, \beta) = -\frac{1}{2N}\sum_{i=1}^{N} w_i(z_i - \beta_0 - x_i^{\mathrm{T}}\beta)^2 + C(\tilde{\beta}_0, \tilde{\beta})^2 \tag{5.55}$$

其中$z_i = \tilde{\beta}_0 + x_i^{\mathrm{T}}\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1-\tilde{p}(x_i))}$。最小化 ℓ_Q 就可得到牛顿更新，这是一个简单的加权最小二乘问题。为了求解正则化问题，可用坐标下降来直接求解式（5.53）。这种方法的缺点是：每维坐标上的最优值不能直接通过公式计算出来，而是需要通过线性搜索得到。根据作者的经验，求解该问题最好是对二次近似采用坐标下降法，这会得到一个迭代算法。对于 λ 的每个值，需要用一个外层循环通过当前参数 $(\tilde{\beta}_0, \tilde{\beta})$ 来计算二次近似 ℓ_Q，然后使用坐标下降来求解带有惩罚项的权重最小二乘问题

$$\underset{(\beta_0, \beta) \in \mathbb{R}^{p+1}}{\text{minimize}} \left\{-\ell_Q(\beta_0, \beta) + \lambda P_\alpha(\beta)\right\} \tag{5.56}$$

将该算法与 5.5.3 节进行比较可知，这是一种广义牛顿算法，求解式（5.56）称为近点牛顿映射（proximal Newton map）详见 Lee et al.（2014）。整个求解过程由下面一系列递归循环构成。

最外层循环：递减 λ。

中间层循环：使用当前参数 $(\tilde{\beta}_0, \tilde{\beta})$ 来计算二次近似 ℓ_Q。

最内层循环：使用坐标下降算法求解带有惩罚项的权重最小二乘问题（5.56）。

当 $p \gg N$ 时，不能一直将 λ 一直减小到零，因为没有定义饱和逻辑斯蒂回归拟合（参数需要取 $\pm\infty$ 才能得到 0 或 1 的概率）。也就是说，这种牛顿算法若不进行步长优化，其收敛性就没有保证（Lee, Lee, Abneel and Ng 2006）。glmnet 包不检查发散性，因为这会减缓计算速度，并且如果按照推荐使用，这种检查不是必须的。作者以启动求解方法给出了一种闭解表达式。并且，从前一次邻近的解中热启动后面的每个求解方法，这通常会使二次近似非常精确。到目前为止，作者采用这种方式并没有碰到过不收敛的情况。

glmnet 包对其他 GLM，比如多分类逻辑斯蒂回归、泊松对数–线性模型以及针对生存数据的 Cox 比例风险模型等，也会用这种求解方法。这些内容在第 3 章给出了详细描述。5.5 节会研究这种方法的效率。

5.5 仿真研究

坐标下降算法和 Nesterov 的组合梯度法对求解 lasso 问题而言都是简单高效的算法。如何比较每次迭代的计算成本？如果当前迭代（给定的某次迭代）的 β^t 有 k 个非零系数，对于 p 个预测子（采用简单更新），坐标下降算法在每个方向上的时间复杂度为 $\mathcal{O}(pN+kN)$。另外，在广义梯度更新（5.31）中，计算矩阵与向量相乘 $\boldsymbol{X}\beta$ 的时间复杂度为 $\mathcal{O}(kN)$，计算 $\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\beta)$ 的时间复杂度为 $\mathcal{O}(pN)$。整个时间复杂度为 $\mathcal{O}(pN+kN)$。

可以通过一个小型仿真实验，来仔细研究坐标下降法、近点梯度下降法以及 Nesterov 动能法之间的相对效率 ①。设训练样本为 $N\times p$ 维矩阵 $\boldsymbol{X}$，特征服从标准高斯分布，而特征之间的相关系数为 0 或 0.5。系数 $|\beta_j|=\exp[-0.5(u(j-1))^2]$，其中 $u=\sqrt{\pi/20}$，其符号会交替地采用 $+1,-1,+1,\ldots$，则 y_i 可由公式

$$y_i=\sum_{j=1}^{p}x_{ij}\beta_j+\sigma\varepsilon_i \tag{5.57}$$

生成，其中 σ 根据信噪比 $\mathrm{Sd}[\mathrm{E}(y_i)]/\sigma=3$ 得到。表 5-1 针对 $N>p$ 和 $N<p$ 两种情形给出了坐标下降法、近点梯度下降法以及 Nesterov 动能法的平均 CPU 时间（标准误差）。对每种情形，正则参数 λ 会取 20 个值，这 20 个值所花的总时间就是该情形下的时间。每种情形都会采用热启动，当参数向量的最大变化小于 10^{-4} 时，就停止迭代。对于后面两种方法，要使用一种近似的回溯线性搜索。从表中可以看到，坐标下降法比其他方法要快 2~6 倍，且对于 $p>N$ 的情况能大幅提高速度。有趣的是，Nesterov 动能法与近点梯度下降法在效率上并不一致，前面的理论也说明了这点。研究表明，是热启动导致了这种不一致，即通过启动就近的解，Nesterov 动能法可以减少“锯齿形”的出现，即便在启动时远离解，该方法也没有太大的问题。

表 5-1　基于 lasso 的线性回归，分别采用坐标下降方法、广义梯度方法和 Nesterov 动能法，在 10 次实现上呈现的 CPU 平均时间（标准误差）。在每一种情形下，所显示的时间都是 20 个 λ 值的路径上的总时间

相关系数	$N=10000,\ p=100$		$N=200,\ p=10000$	
	0	0.5	0	0.5
坐标下降	0.110 (0.001)	0.127 (0.002)	0.298 (0.003)	0.513 (0.014)
近点梯度	0.218 (0.008)	0.671 (0.007)	1.207 (0.026)	2.912 (0.167)
Nesterov	0.251 (0.007)	0.604 (0.011)	1.555 (0.049)	2.914 (0.119)

① 感谢 Jerome Friedman 为本节提供程序。

表 5-2 给出了基于逻辑斯蒂回归的结果。这里也用前面的方法来生成训练数据，但有 15 个符号交替变化的非零的 β_j，$|\beta_j| = 15 - j + 1$。令 $p_i = 1/(1 + \exp(-\sum x_{ij}\beta_j))$，$y_i$ 的取值为 0 或 1，概率为 $\text{Prob}\,(y_i = 1) = p_i$。

表 5-2　基于 lasso 的逻辑斯蒂回归，分别采用坐标下降方法、广义梯度方法和 Nesterov 动能法，在 10 次实现上呈现的 CPU 平均时间（标准误差）。在每一种情形下，所显示的时间都是 20 个 λ 值的路径上的总时间

相关系数	$N = 10000, p = 100$		$N = 200, p = 10000$	
	0	0.5	0	0.5
坐标下降	0.309 (0.086)	0.306 (0.086)	0.646 (0.006)	0.882 (0.026)
近点梯度	2.023 (0.018)	6.955 (0.090)	2.434 (0.095)	4.350 (0.133)
Nesterov	1.482 (0.020)	2.867 (0.045)	2.910 (0.106)	8.292 (0.480)

从表 5-2 可以看出，坐标下降法比其他方法要快 5~10 倍，且对于 $p > N$ 的情况能大幅提高速度。与表 5-1 一样，Nesterov 动能法与近点梯度下降法在效率上并不一致。

读者也可以像上面那样，从细微方面着手比较这些算法之间的差异，因为这些方法的性能与实现细节有关。也许读者会有更进一步的疑问，因为这里表现最好的算法（坐标下降法）是由本书的两位作者联合提出的。因此，我们只能说已经尽量公平地对待所有方法了，也尽量让所编写的代码有效。更重要的是，为了让读者进一步研究，所有用来产生结果的脚本和程序都可在本书的网站下载。

5.6　最小角回归

最小角回归（LAR）是一种**同伦**（homotopy）算法，可以用来求解 lasso 问题。该算法会将整个解的路径作为正则参数 λ 的一个函数。最小角回归是一种有效的算法，但它并不像本章其他一些方法那样适合大规模问题。它的统计学动机很有意义，可将其看成是向前逐步回归的民主（democratic）版。

向前逐步回归通过每次增加一个变量来建立模型，即每步都会挑出最好的变量放到**活动集**中，然后用活动集中所有的变量来进行最小二乘拟和。最小角回归采用了类似的策略，但只会尽可能多地输入适合模型的变量。第一步得到与输出最相关的变量。LAR 方法会不断朝最小二乘值移动该变量的系数（这会让它与残差的相关性在绝对值上减少），而不是拟和该变量。一旦别的变量与残差的相关性达到该值，这个过程就会暂停，并将第二个变量加到活动集中，按相同方式修改它们的系数，并将相关系数绑定，共同减小。当模型中的所有变量都进行了一次最小二乘拟和后，这个过程才会停止。算法 5.1 详细描述了最小角回归算法。虽然是按相关

性来介绍 LAR 的，但由于特征已经进行归一化，因此用内积来得到相关性比较容易。需要解释一下该算法第 3 步中的 K：如果 $p > N-1$，在经过 $N-1$（-1 来自模型中的截距，我们通过中心化数据解决这种问题）步之后，LAR 会得到残差为零的解。

算法 5.1 最小角回归

1. 对训练样本进行归一化，使其样本匀值为零，ℓ_2 范数为 1。初始残差为 $\boldsymbol{r}_0=\boldsymbol{y}-\bar{\boldsymbol{y}}$，$\beta^0=(\beta_1,\beta_2,\ldots,\beta_p)=\mathbf{0}$。
2. 找到与 $\boldsymbol{r}_0$ 最相关的 $\boldsymbol{x}_j$，即 $\lambda_0=\max_j|\langle\boldsymbol{x}_j,\boldsymbol{r}_0\rangle|$，定义活动集 $\mathcal{A}=\{j\}$ 和一个由单变量构成的矩阵 $\boldsymbol{X}_{\mathcal{A}}$。
3. 对于 $k=1,2,\ldots,K=\min(N-1,p)$，进行以下操作：
   (a) 定义最小二乘方向 $\delta=\dfrac{1}{\lambda_{k-1}}\left(\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}}\right)^{-1}\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{r}_{k-1}$，对于 p 维向量 Δ，使 $\Delta_{\mathcal{A}}=\delta$，其他元素都为 0。
   (b) 以 Δ 方向，从 β^{k-1} 开始，朝着它们的最小二乘解移动系数 β，其中 $\boldsymbol{X}_{\mathcal{A}}$: $\beta(\lambda)=\beta^{k-1}+(\lambda_{k-1}-\lambda)\Delta$，$0\leqslant\lambda\leqslant\lambda_{k-1}$。进一步得到新的残差 $\boldsymbol{r}(\lambda)=\boldsymbol{y}-\boldsymbol{X}\beta(\lambda)=\boldsymbol{r}_{k-1}-(\lambda_{k-1}-\lambda)\boldsymbol{X}\Delta$。
   (c) 关注 $\lambda=|\langle\boldsymbol{x}_\ell,\boldsymbol{r}(\lambda)\rangle|$，其中 $\ell\notin\mathcal{A}$，设第 j 个变量让 λ 达到最大，即 $|\langle\mathbf{x}_j,r(\lambda)\rangle|=\lambda$，则应该将 j 加入 $\mathcal{A}$，且令 $\lambda_k=\lambda$。
   (d) 让 $\mathcal{A}=\mathcal{A}\cup\{j\}$，$\beta^k=\beta(\lambda_k)=\beta^{k-1}+(\lambda_{k-1}-\lambda_k)\Delta$，$\boldsymbol{r}_k=\boldsymbol{y}-\boldsymbol{X}\beta^k$。
4. 返回序列 $\left\{\lambda_k,\beta^k\right\}_0^K$。

下面解释这个算法。在步骤 3b 中，很容易得到 $|\langle\boldsymbol{x}_j,r(\lambda)\rangle|=\lambda(\forall j\in\mathcal{A})$，即相关系数会同这个路径有紧密联系，并且会随 λ 的降低而减少至 0。事实上，$\beta^0=\beta^{k-1}+\lambda_{k-1}\Delta$ 是对应 $\mathcal{A}$ 的最小二乘系数向量。

通过 LAR 来构造系数与分段线性方式不一样。图 5-9 的左图为 LAR 系数轮廓，这会演化成 ℓ_1 弧长的函数 ①。注意，在步骤 3c 中不需要取一个小的步长并重新检查相关性。变量 ℓ 赶上（catching up）是指 $|\langle\boldsymbol{x}_\ell,\boldsymbol{r}(\lambda)\rangle|=\lambda$，这是一对关于 λ 的线性方程。对于所有 $\ell\notin\mathcal{A}$，都会求解方程得到相应的 λ，最大 λ 对应的 ℓ 变量会加到 $\mathcal{A}$ 中（见习题 5.9）。

图 5-9 右图是 lasso 在相同数据下得到的系数轮廓。左右两图基本一样，第一次不同存在于粉红色系数线穿越水平轴时。基于这样的观察，对 LAR 算法步骤 3c 要进行一个小的修改，即让这一步也为分段线性。

3c+lasso 修改：如果在下一个变量被放入 $\mathcal{A}$ 之前，非零系数穿过了零点，就将对应的变量从 $\mathcal{A}$ 中删除，并重新计算当前的联合最小二乘方向。

注意，图中的粉红色系数路径会有一段一直为零，然后再次变为有效，但这次变成了负值。

① 可微曲线 $\{s\mapsto\beta(s)|s\in[0,S]\}$ 的 ℓ_1 弧长可由 $TV(\beta,S)=\int_0^S\|\dot{\beta}(s)\|_1\mathrm{d}s$ 来得到，其中 $\dot{\beta}(s)=\partial\beta(s)/\partial s$。基于分段线性的 LAR 的系数轮廓，这相当于在迭代过程中前后两步有变化的系数的 ℓ_1 范数加在一起。

这两种算法得到如此相似的结果是有原因的。观察可知，在算法中的任何阶段都有

$$\boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\beta(\lambda))=\lambda\cdot s_j,\quad \forall j\in\mathcal{A} \tag{5.58}$$

其中 $s_j\in\{-1,1\}$ 表示内积的符号，λ 在方程左右两边都有。由 LAR 活动集的定义可知：$\left|\boldsymbol{x}_k^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\beta(\lambda))\right|\leqslant\lambda$，$\forall k\notin\mathcal{A}$。lasso 的目标函数为 ①

$$R(\beta)=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2+\lambda\|\beta\|_1 \tag{5.59}$$

在给定 λ 的情况下，令 $\mathcal{B}$ 为解中变量的活动集。对这些变量而言，$R(\beta)$ 是可微的，其稳定条件为

$$\boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\beta)=\lambda\cdot\operatorname{sgn}(\beta_j),\qquad \forall j\in\mathcal{B} \tag{5.60}$$

将式（5.60）与式（5.58）进行比较会发现，仅当 β_j 的符号与内积的符号匹配时它们是一样的。这就是活动系数穿过零点时，LAR 和 lasso 出现差别的原因。在步骤 3c+ 中，若有变量不符合式（5.60），则将其从活动集 $\mathcal{B}$ 中删除。习题 5.9 表明，在 LAR 的更新过程中，当 λ 减少时，等式意味着系数分段线性特性。对于非活动变量，稳定条件为

$$|\boldsymbol{x}_k^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\beta)|\leqslant\lambda,\qquad \forall k\notin\mathcal{B} \tag{5.61}$$

这与 LAR 算法一致。

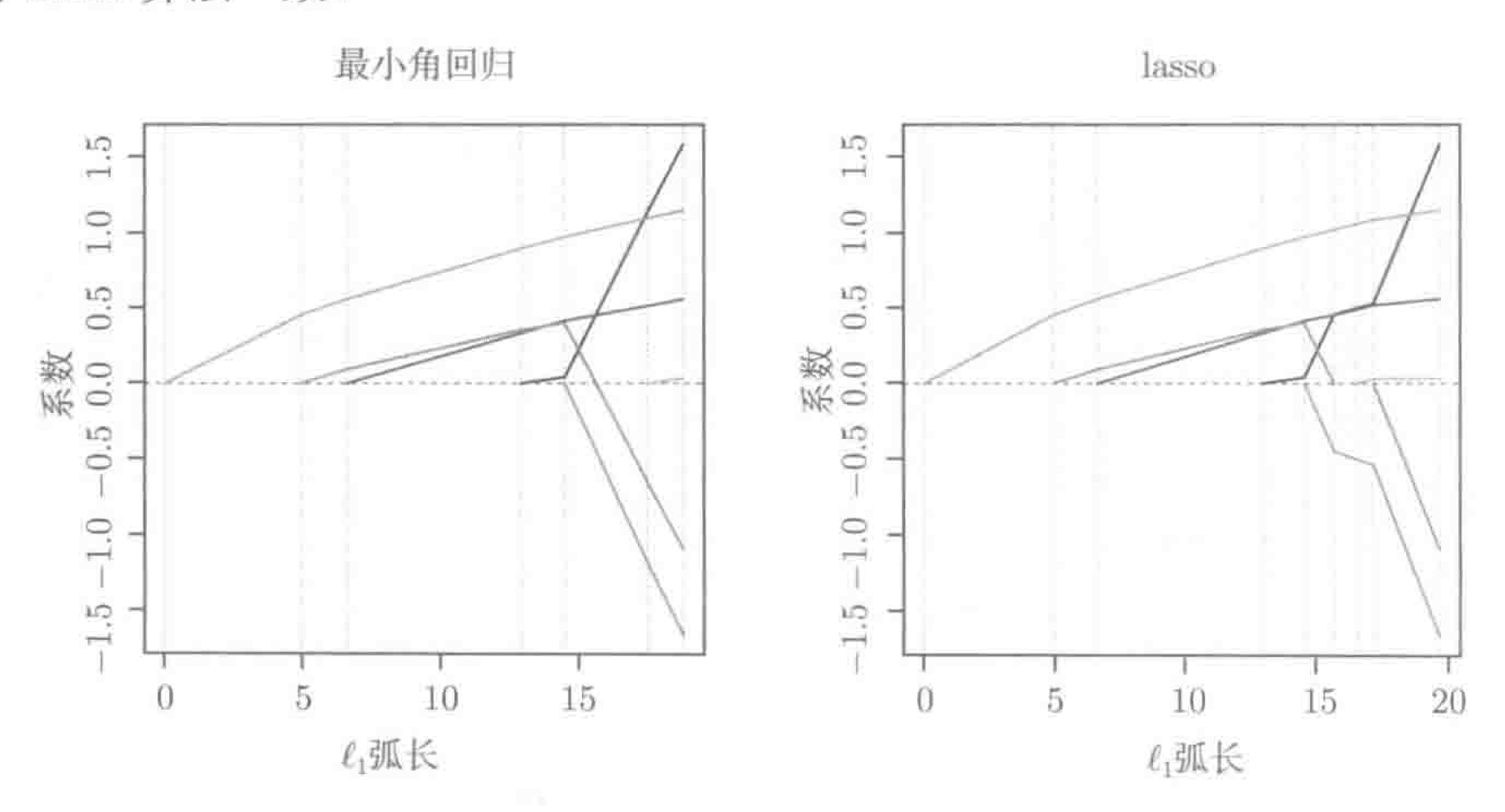


图 5-9 左图为 LAR 在模拟数据上得到的系数轮廓，这些系数与 L^1 弧长之间有函数关系。右图是求解 lasso 问题所得到的系数轮廓。二者在红色系数轮廓穿过水平轴（弧长为 16 附近）之前是一样的（见彩插）

LAR 算法利用了 lasso 系数路径是分段线性的事实。这个性质对更一般的问题也成立，详见 Rosset and Zhu（2007）。

① 这里去掉了 $\frac{1}{N}$，以便与 LAR 求解过程相吻合。因此，λ 的所有值比真实值要大 N 倍。

5.7 交替方向乘子法

交替方向乘子法（Alternating Direction Method Of Multiplier，ADMM）属于拉格朗日方法，对大规模问题，它有很多吸引人的性质。交替方向乘子法吸收了多种经过长时间发展的方法的思想。这里只对该算法进行简单概述，读者可参考 Boyd et al.（2011）来全面了解该算法。

假设要求解的优化问题为

$$\underset{\beta\in\mathbb{R}^m,\theta\in\mathbb{R}^n}{\text{minimize}}\ f(\beta)+g(\theta)\ ,\quad \boldsymbol{A}\beta+\boldsymbol{B}\theta=c \tag{5.62}$$

其中 $f:\mathbb{R}^m\to\mathbb{R}$，$g:\mathbb{R}^n\to\mathbb{R}$ 都是凸函数。$\boldsymbol{A}\in\mathbb{R}^{n\times d}$，$\boldsymbol{B}\in\mathbb{R}^{n\times d}$ 都是约束矩阵，$c\in\mathbb{R}^d$ 为约束向量。为了求解该问题，需引入与约束相关的拉格朗日乘子，这些乘子构成向量 $\mu\in\mathbb{R}^d$。然后考虑增广拉格朗日函数

$$L_\rho(\beta,\theta,\mu):=f(\beta)+g(\theta)+\langle\mu,\boldsymbol{A}\beta+\boldsymbol{B}\theta-c\rangle+\frac{\rho}{2}\|\boldsymbol{A}\beta+\boldsymbol{B}\theta-c\|_2^2 \tag{5.63}$$

其中 $\rho>0$ 是一个很小的固定参数。与 ρ 相关的二次项是一个增广拉格朗日项，它以较光滑的方式被强制靠近约束条件。ADMM 算法会不断分别最小化以 β 和 θ 为变量的增广拉格朗日函数（5.63），然后采用对偶变量更新 μ。更新公式为

$$\beta^{t+1}=\underset{\beta\in\mathbb{R}^m}{\text{arm}\min}\,L_\rho(\beta,\theta^t,\mu^t) \tag{5.64a}$$

$$\theta^{t+1}=\underset{\theta\in\mathbb{R}^n}{\text{arm}\min}\,L_\rho(\beta^{t+1},\theta,\mu^t) \tag{5.64b}$$

$$\mu^{t+1}=\mu^t+\rho(\boldsymbol{A}\beta^{t+1}+\boldsymbol{B}\theta^{t+1}-c) \tag{5.64c}$$

在此通过 $t=0,1,2,\ldots$ 进行迭代。式（5.64c）会对拉格朗日乘子向量 μ 进行更新，这属于对偶上升更新。在相对宽松的条件下，这个迭代过程收敛于问题（5.62）的最优解。

ADMM 的框架有几个优势。第一，可将带不可微约束的凸问题的参数分离成 β 和 θ，从而使问题很容易处理。下面会通过 lasso 来解释这个迭代过程。第二，ADMM 可将大规模问题分成较小的片。有大量观测值的数据集可将数据分成较小的块，并对每个块进行优化。习题 5.12 会进行更详细的讨论。每个数据块上都应该有约束条件，以确保优化每个数据块而得到的解向量与其他收敛的方法一致。同理，可以将问题按特征块分解，并用分块坐标回溯方式求解。

例 5.8：用 ADMM 求解 lasso 问题　lasso 的拉格朗日形式可等价表示成

$$\underset{\beta\in\mathbb{R}^p,\theta\in\mathbb{R}^p}{\text{minimize}}\left\{\frac{1}{2N}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2+\frac{1}{N}\lambda\|\theta\|_1\right\},\quad \beta-\theta=0 \tag{5.65}$$

对于 lasso，ADMM 的更新公式为

$$
\begin{aligned}
\beta^{t+1} &= (\boldsymbol{X}^T\boldsymbol{X} + \rho\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{y} + \rho\theta^t - \mu^t) \\
\theta^{t+1} &= \mathcal{S}_{\lambda/\rho}(\beta^{t+1} + \mu^t/\rho), \\
\mu^{t+1} &= \mu^t + \rho(\beta^t - \theta^{t+1}).
\end{aligned}
\tag{5.66}
$$

从上式可看出，对 β 的更新是岭回归，对 θ 的更新是软阈值，对 μ 采用线性更新即可。在这三个更新中，第一个更新的计算量最大，在最初完成 $\boldsymbol{X}$ 的奇异值分解后，接下来的一系列迭代就会很快了。$\boldsymbol{X}$ 的奇异值分解的时间复杂度为 $\mathcal{O}(p^3)$，但可以事先计算好。后面一系列迭代的时间复杂度为 $\mathcal{O}(Np)$。因此，在初始化后，每次迭代的时间复杂度与坐标下降法或者组合梯度法差不多。

5.8　优化–最小化算法

本节将介绍优化–最小化（Majorization-Minimization 或 Minorization-Maximization，MM）算法，这种方法对非凸问题尤其有效。这些算法属于辅助变量法，因为它们都基于引入的额外变量来优化目标函数，使其最小化。虽然这些方法广泛用于带约束的问题，但这里将介绍它们在简单的无约束问题上的应用，这类无约束问题的形式为 $\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\, f(\beta)$，其中 $f:\mathbb{R}^p \mapsto \mathbb{R}$（可能）是一个非凸函数。

如果条件

$$
f(\beta) \leqslant \Psi(\beta,\theta), \quad \theta \in \mathbb{R}^p \tag{5.67}
$$

成立，则函数 $\Psi:\mathbb{R}^p\times\mathbb{R}^p \mapsto \mathbb{R}$ 在点 $\beta\in\mathbb{R}^p$ 处优化函数 f。当 $\beta=\theta$ 时，等号成立 [显然，minorization 的定义就是将式（5.67）中的 $\leqslant$ 换成 $\geqslant$]。图 5-10 给出了优化（majorizing）函数的示意图。

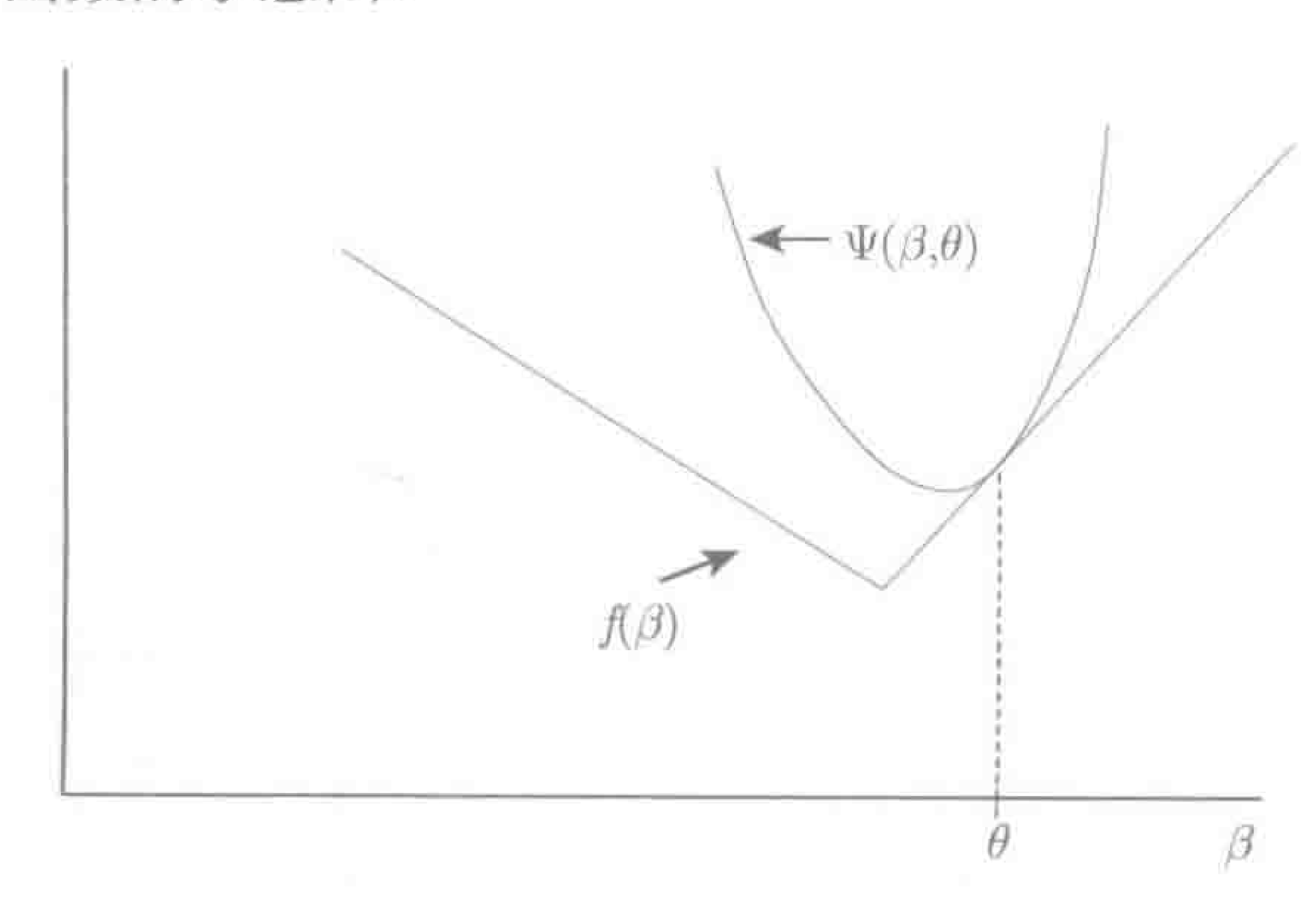


图 5-10　MM 算法中优化函数的示意图。函数 $\Psi(\beta,\theta)$ 位于 $f(\beta)$ 之上。当 $\beta=\theta$ 时，$f(\beta)=\Psi(\beta,\theta)$，MM 算法通过求解一系列与函数 Ψ 相关的子问题得到目标函数 f 的最小值

MM 算法从初始值 β_0 开始，通过公式

$$\beta^{t+1} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \Psi(\beta, \beta^t), \qquad t = 0, 1, 2, \ldots \tag{5.68}$$

进行迭代更新，从而得到目标函数 f 的最小值。由式（5.67）的性质可知，这种方式可以生成一组序列，使得目标函数 $f(\beta^t)$ 非递增，即

$$f(\beta^t) = \Psi(\beta^t, \beta^t) \overset{\text{(i)}}{\geqslant} \Psi(\beta^{t+1}, \beta^t) \overset{\text{(ii)}}{\geqslant} f(\beta^{t+1}) \tag{5.69}$$

不等式 (i) 成立是因为 β^{t+1} 是函数 $\beta \mapsto \Psi(\beta, \beta^t)$ 的最小值，不等式 (ii) 成立是因为式（5.67）的性质。如果 f 是严格的凸函数，则 MM 算法会收敛到全局最优解。

不同的问题，会使用不同类型的优化函数。通常一个好的优化函数要能相对容易地计算更新式（5.68），至少要比直接得到 f 的最小值要容易一些。更详细的介绍可参见 Lange（2004）。

例 5.9: 近点梯度法可看成 MM 算法 5.3.3 节的近点梯度法能运用于可分解为 $f = g + h$ 形式的目标函数，其中 g 为可微凸函数，h 为凸函数但可能不可微。对 g 采用二阶泰勒级数展开，可得到

$$\begin{aligned} f(\beta) &= g(\beta) + h(\beta) \\ &= g(\theta) + \langle \nabla g(\theta), \theta - \beta \rangle + \frac{1}{2} \left\langle \theta - \beta, \nabla^2 g(\beta')(\theta - \beta) \right\rangle + h(\beta) \end{aligned}$$

其中 $\beta' = s\beta + (1 - s)\theta$，$s \in [0, 1]$。由于梯度 ∇g 满足 Lipschiz 条件（5.27），这说明海森矩阵有一致上界，即 $\nabla^2 g(\beta') \prec L\boldsymbol{I}_{p \times p}$，因此有

$$f(\beta) \leqslant \underbrace{g(\theta) + \langle \nabla g(\theta), \theta - \beta \rangle + \frac{L}{2} \|\theta - \beta\|_2^2 + h(\beta)}_{\Psi(\beta, \theta)}$$

在这个式子中，当 $\beta = \theta$ 时等号成立。因此，在选择具体的优化函数后，近点梯度法可看成是 MM 算法。

除了直接使用海森的上界外，还有其他方法可以得到优化函数，比如：Jensen 不等式可用于推导常见的 EM 算法，这就属于 MM 算法（Hunter and Lange 2004, Wu and Lange 2007）。第 8 章还会证明 MM 算法对稀疏多变量分析很有用。

5.9 双凸问题和交替最小化

5.4 节的坐标下降法也适用于优化一类非凸函数，即双凸函数。对于函数 $f: \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$，若对任意的 $\beta \in \mathbb{R}^n$，则函数 $\alpha \mapsto f(\alpha, \beta)$ 是凸函数；若对任意的 $\alpha \in \mathbb{R}^m$，函数 $\beta \mapsto f(\alpha, \beta)$ 是凸函数。显然，基于 (α, β) 对的联合凸函数一定是双凸函数，但双凸函数却不是联合凸函数。例如，考虑双凸函数

$$f(\alpha, \beta) = (1 - \alpha\beta)^2, \quad |\alpha| \leqslant 2, \quad |\beta| \leqslant 2 \tag{5.70}$$

如图 5-11 所示，沿着横轴方向和纵轴方向所得的切片为凸函数，但按其他方向得到的切片为非凸函数。

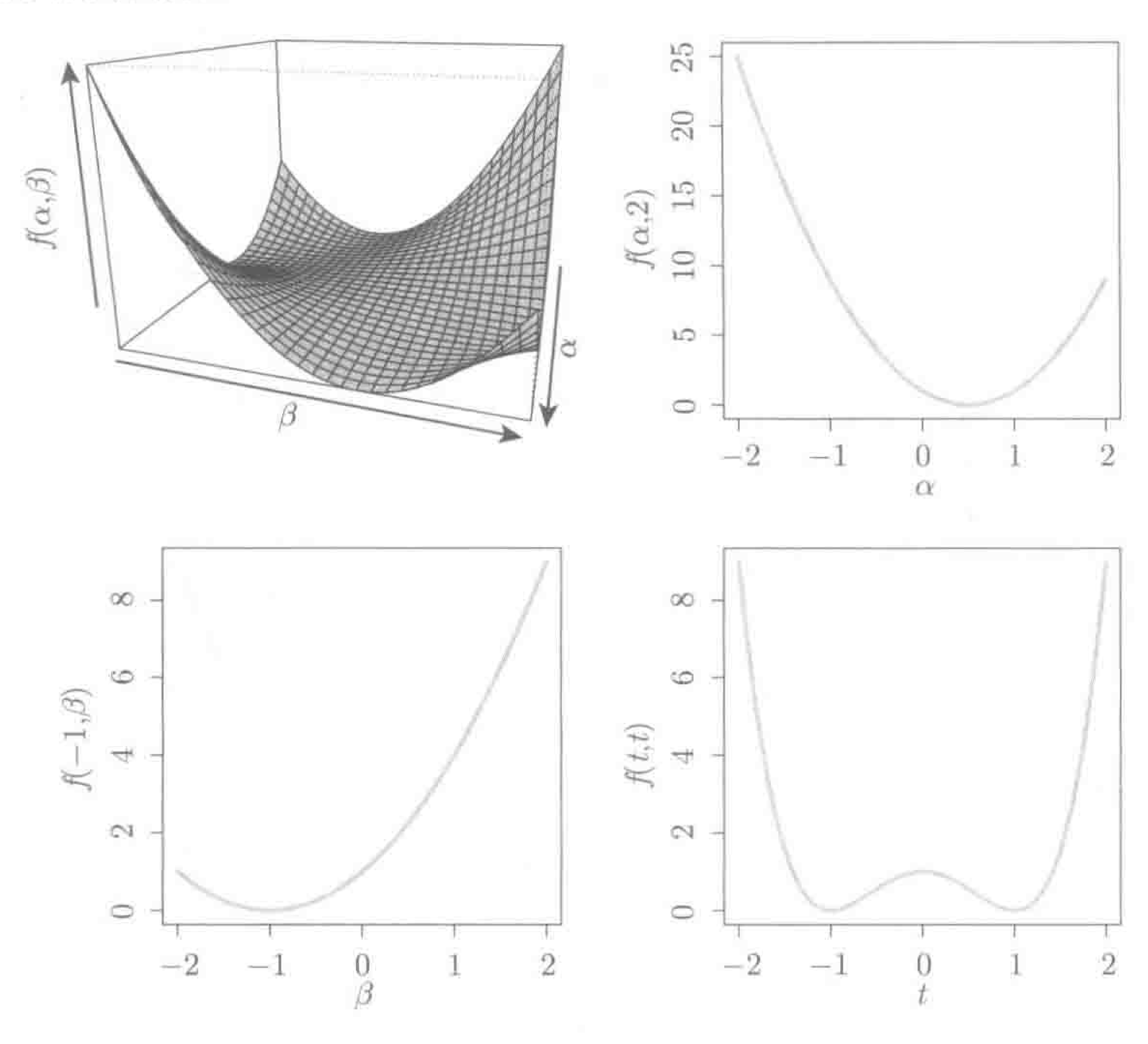


图 5-11　双凸函数的例子。左上图为函数 $f(\alpha, \beta) = (1-\alpha\beta)^2$，其中 $-2 \leqslant \alpha \leqslant 2$，$-2 \leqslant \beta \leqslant 2$。右上图和左下图分别为 $\beta = 2$，$\alpha = -1$ 时 $f(\beta, \alpha)$ 的图形；右下图为 $\alpha = \beta = t$ 时 $f(\beta, \alpha)$ 的图形

更一般而言，设 $\mathcal{A} \subseteq \mathbb{R}^m$，$\mathcal{B} \subseteq \mathbb{R}^n$ 是两个非空的凸集，$\mathcal{C} \subseteq \mathcal{A} \times \mathcal{B}$。对于每个给定的 $\alpha \in \mathcal{A}$ 和 $\beta \in \mathcal{B}$，

$$\mathcal{C}_\alpha := \{\beta \in \mathcal{B} | (\alpha, \beta) \in \mathcal{C}\} \quad \text{和} \quad \mathcal{C}_\beta := \{\alpha \in \mathcal{A} | (\alpha, \beta) \in \mathcal{C}\} \tag{5.71}$$

两个集合分别称为集合 $\mathcal{C}$ 的 α 部分和 β 部分。如果对于每个 $\alpha \in \mathcal{A}$，$\mathcal{C}_\alpha$ 为凸集，对于每个 $\beta \in \mathcal{B}$，$\mathcal{C}_\beta$ 为凸集，则 $\mathcal{C} \subseteq \mathcal{A} \times \mathcal{B}$ 称为**双凸集**。给定一个双凸集 $\mathcal{C}$，若对任意固定的 $\beta \in \mathcal{B}$，函数 $\alpha \mapsto f(\alpha, \beta)$ 是凸函数，对任意固定的 $\alpha \in \mathcal{A}$，函数 $\beta \mapsto f(\alpha, \beta)$ 是凸函数，则函数 $f: \mathcal{C} \to \mathbb{R}$ 为**双凸函数**。

基于上面的介绍可知，双凸优化问题具有形式 $\underset{(\alpha,\beta) \in C}{\text{minimize}} f(\alpha, \beta)$，$\mathcal{C}$ 为 $\mathcal{A} \times \mathcal{B}$ 上的双凸集，目标函数在 $\mathcal{C}$ 上为双凸函数。

求解双凸优化问题最著名的方法是交替凸搜索法（Alternate Convex Search，ACS），这种方法会对 α 和 β 块采用块坐标下降法：

(a) 用 $\mathcal{C}$ 中的一些点作为初始化的 (α^0,β^0)；

(b) 迭代 $t=0,1,2,\ldots$

① 固定 $\beta=\beta^t$，然后执行 $\alpha^{t+1}\text{argmin}_{\alpha\in\mathcal{C}_{\beta^t}} f(\alpha,\beta^t)$；

② 固定 $\alpha=\alpha^{t+1}$，然后执行 $\beta^{t+1}\text{argmin}_{\beta\in\mathcal{C}_{\alpha^{t+1}}} f(\alpha^{t+1},\beta)$。

对于给定的双凸结构，这两个更新会求解凸优化问题。若能较快求解这两个凸优化子问题，则 ACS 算法还是很有效的。

这种构造的函数值序列 $\{f(\alpha^t,\beta^t)\}_{t=0}^{\infty}$ 为非递增序列。如果 f 在 $\mathcal{C}$ 上为有界函数，则函数值序列会收敛到某个极限值。注意，这种形式的收敛相对较弱，而且只能确保函数值收敛，解的序列 $\{(\alpha^t,\beta^t)\}$ 可能不收敛，在某些情形下会发散至无穷。假设解序列收敛，那会收敛到什么地方呢？由于双凸函数 f 整体并不凸，因此不一定会收敛到全局最优解。所以在一般情况下，也只是收敛到局部最优解。

更具体而言，当满足条件

$$f(\alpha^*,\beta^*)\leqslant f(\alpha^*,\beta),\quad \beta\in\mathcal{C}_{\boldsymbol{\alpha}^*}$$

$$f(\alpha^*,\beta^*)\leqslant f(\alpha,\beta^*),\quad \beta\in\mathcal{C}_{\boldsymbol{\beta}^*}$$

$(\alpha^*,\beta^*)\in\mathcal{C}$ 是局部最优解。

例 5.10: 交替子空间算法 用来求解双凸问题的 ACS 算法的收敛可用交替子空间算法来描述，交替子空间算法可用来计算矩阵的最大奇异向量/奇异值。给定矩阵 $\boldsymbol{X}\in\mathbb{R}^{m\times n}$，找出该矩阵在 Frobenius 范数下最好的秩 1 近似 ①。这个近似问题会最小化目标函数

$$f(\alpha,\beta,s)=\left\|\boldsymbol{X}-s\alpha\beta^{\mathrm{T}}\right\|_F^2 \tag{5.72}$$

其中 $\alpha\in\mathbb{R}^m$，$\beta\in\mathbb{R}^n$，而且 $\|\alpha\|_2=\|\beta\|_2=1$，$s>0$ 是一个标量。ACS 算法可以从任意一个随机单位向量 β^0 开始，然后按公式（$t=0,1,2,\ldots$）

$$\alpha^t=\frac{\boldsymbol{X}\beta^{t-1}}{\|\boldsymbol{X}\beta^{t-1}\|_2}\ ,\ \beta^t=\frac{\boldsymbol{X}^{\mathrm{T}}\alpha^{t-1}}{\left\|\boldsymbol{X}^{\mathrm{T}}\alpha^{t-1}\right\|_2} \tag{5.73}$$

进行迭代更新。在收敛处，标量 s 可由 $s=\|\boldsymbol{X}\beta^t\|_2$ 计算得到。可证明（见习题 5.3）：只要 β^0 不与最大的右奇异向量正交，迭代 (α^t,β^t) 会分别会收敛到 $\boldsymbol{X}$ 的最大奇异值所对应的左奇异向量和右奇异向量。

这个过程与用来求解对称半正定矩阵的最大特征向量的 power 方法有关。对应右奇异向量的 β^t 的迭代公式为

$$\beta^{t+1}=\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\beta^t}{\left\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\beta^t\right\|_2} \tag{5.74}$$

① 矩阵的 Frobenius 范数是将欧式范数应用到了将该矩阵向量化后的版本。

α^t 的情况与之类似。因此，这个过程能加速算子 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$，这种归一化会让最大特征值之外的所有值都为零。关于 power 方法更详细的描述参见 De Leeuw (1994) 以及 Golub and Loan (1996, §7.3)。

第 7 章的算法 7.2 是 ACS 算法的另一个例子。

5.10 筛选规则

由 5.6 节可知，内积在 lasso 问题上起着重要作用。为了简洁，在此假设所有变量都进行了中心化（这样可以不用考虑截距）。我们要求解的 lasso 问题[①]是

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ \ \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{X}\beta\right\|_2^2+\lambda\left\|\beta\right\|_1 \tag{5.75}$$

其中 λ 的值为递减序列。模型的第一个变量有最大内积绝对值 $\lambda_{\max}=\max\limits_j|\langle\boldsymbol{x}_j,\boldsymbol{y}\rangle|$，这也定义了 λ 的初始值。在整个迭代过程中，活动集中的任意一个元素 $\boldsymbol{x}_j$ 都有 $|\langle\boldsymbol{x}_j,\boldsymbol{y}-\hat{\boldsymbol{y}}_\lambda\rangle|=\lambda$，与残差有较小内积的元素会被淘汰掉。因此可以认为，有较小内积的元素相比有较大内积的元素更有可能对应系数 0。基于这样的观点，可以去掉一些特征，从而提高计算效率。比如在基因应用中，可能有成上百万的变量 SNP，但只有少数变量对拟和模型有用。本节将介绍基于这种思想的筛选规则，采用这种规则后，不但可以加快计算，而且仍能得到精确解。

下面介绍对偶多面体投影（Dual Polytope Projection，DPP）规则（Wang, Lin, Gong, Wonka and Ye 2013）。假设在 $\lambda<\lambda_{\max}$ 下计算 lasso 的解。如果第 j 个变量满足条件

$$\left|\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}\right|<\lambda_{\max}-\left\|\boldsymbol{x}_j\right\|_2\left\|\boldsymbol{y}\right\|_2\frac{\lambda_{\max}-\lambda}{\lambda} \tag{5.76}$$

则可根据 DPP 规则删除它。也许读者会惊讶于这样的规则会有用，其实作者第一次看到它时也感到惊讶。在线性回归中，一个特征[②]可能会显得微不足道，但当它与其他特征一起用于模型中时，就会起到很大的作用。同样的现象也出现在 lasso 中。

事实上，这里并不矛盾，类似的规则可应用于正则化路径的任何阶段（不仅仅是开端）。假设在 λ' 处有 lasso 的解 $\beta(\lambda')$，需要在 $\lambda<\lambda'$ 的地方筛选掉一些变量。那么若

$$\left|\boldsymbol{x}_j^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}(\lambda')\right)\right|<\lambda'-\left\|\boldsymbol{x}_j\right\|_2\left\|\boldsymbol{y}\right\|_2\frac{\lambda'-\lambda}{\lambda} \tag{5.77}$$

则第 j 个变量不是在 λ 处的活动集的一部分。本书称这种规则为序列化的 DPP 规则。

① 本节去掉了 lasso 公式第一部分中的 $\frac{1}{N}$（同参考文献中的公式一致），因此 λ 的所有值比真实值要大 N 倍。

② 也称预测子数。——译者注

从图 5-12 可看出，在模拟训练样本集（有 5000 个特征，图中有详细说明）上，这个规则的性能如何。对于所有 λ 的值，采用全局 DPP 规则（5.76）后，所有特征很快都被选了出来。在 λ 较小时，只有 8 个特征进入模型，即便此时所有 5000 个特征都选了出来。但序列 DPP 规则的情况就要好得多，在模型中有 250 个特征时，所选择的特征数量为 1200 个。因此，若 λ' 和 λ 靠近，则序列筛选规则（5.77）有较好的性能。附录 B 会推导出 lasso 对偶和 DPP 规则。

为了得到更好的性能，可以考虑改进 DPP 规则，使其不那么保守，而且允许偶尔错误。这种改进的规则是整个更新策略的一部分，能在迭代收敛时得到精确的解。改进的全局 DPP 规则（5.76）称为**全局强规则**（global strong rule），即当

$$\left|\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}\right| < \lambda - (\lambda_{\max} - \lambda) = 2\lambda - \lambda_{\max} \tag{5.78}$$

时丢掉特征 j。与全局 DPP 规则相比，全局强规则会丢掉更多的特征（见图 5-12 蓝色的点与橙色的点）。同样，如果满足条件

$$\left|\boldsymbol{x}_j^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda')\right)\right| < 2\lambda - \lambda' \tag{5.79}$$

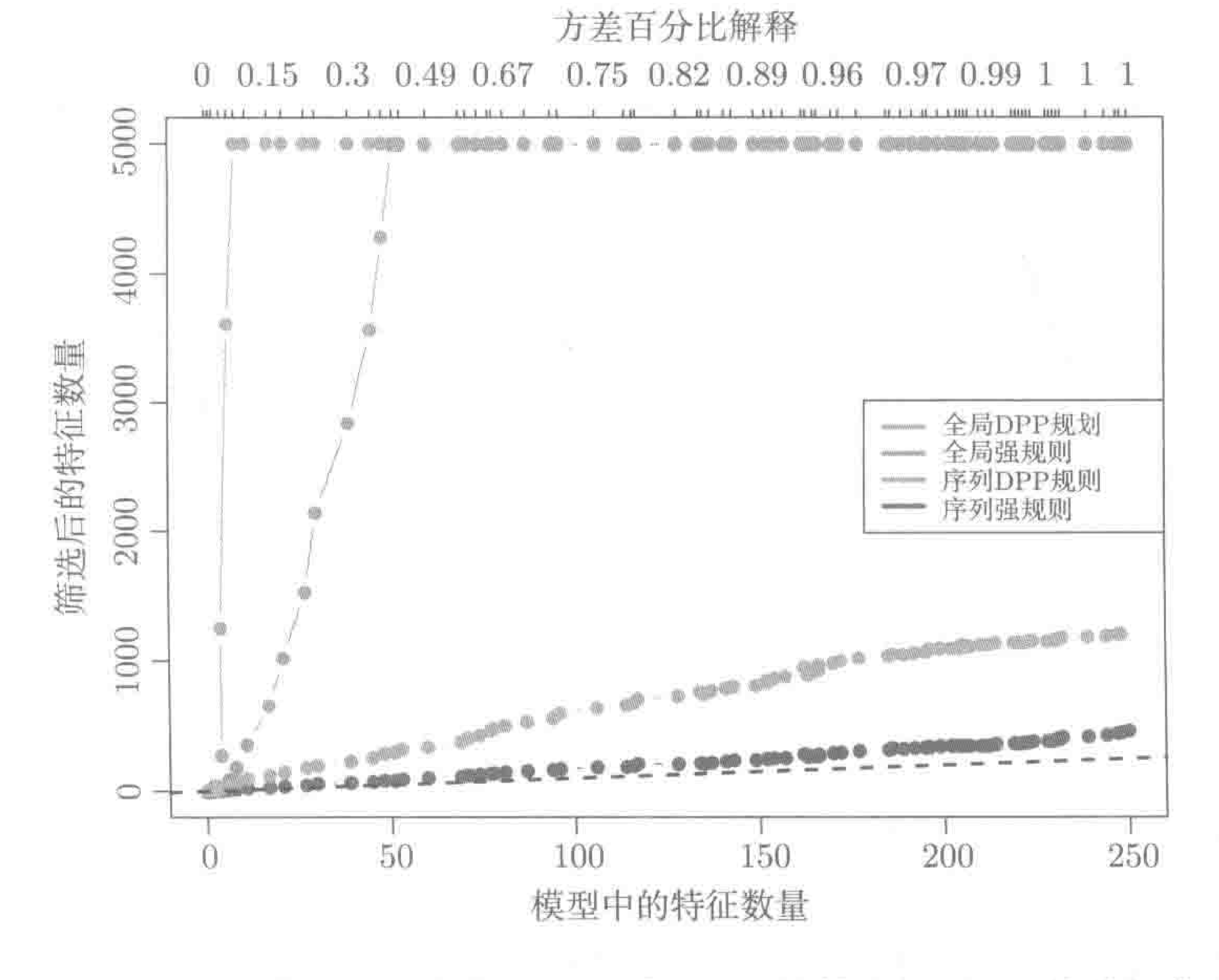


图 5-12 lasso 回归：在模拟的训练数据集上采用不同的筛选规则。训练数据集总共有 $N=200$ 个样本，每个样本特征 $p=5000$。这些特征是由不相关的高斯分布生成的。真实的系数中有四分之一不为零。图中显示了每次筛选出来的一部分特征，以及在给定 λ 的情况下模型中的特征数量。λ 的值从左到右依次减少。在图中，λ 总共取了 100 个值，这些值按对数尺度均匀分开。为了便于参考，图中还增加了一根虚线，斜率为 1。在图的顶部给出了模型的方差比百分解释。这两种强规则并不冲突

序列强规则也会从优化问题丢掉在 λ 处的第 j 特征。简而言之，活动集中包含的特征能够通过与残差求内积得到 λ，所以可将那些在 $\lambda' > \lambda$ 处，与当前残差的内积靠近 λ 的特征加进来，其靠近程度由 $\lambda' - \lambda$ 决定。

与序列 DPP 规则一样，序列强规则也通过一系列递减的 λ 来求解 lasso 问题。图 5-12 给出了全局和序列的强规则它们的性能都要比与之相对应的一般规则好。在这个例子中，所有强规则都没有犯任何错误。这里所说的错误是指在求解过程中丢掉了非零系数对应的特征。序列强规则（5.79）有非常好的性能，丢掉了大多数冗余的特征。

下面对强规则做进一步解释（Tibshirani, Bien, Friedman, Hastie, Simon, Taylor and Tibshirani$_2$ 2012）。假设在 $\lambda = \lambda_{\max}$ 时，第 j 个特征不在模型中，则 lasso 的 KKT 条件可确保有 $|\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}| < \lambda_{\max}$，因此，可将全局规则（5.78）解释为从 $\lambda_{\max}$ 移动 λ 时，内积 $|\boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}(\lambda))|$ 最多可增多 $\lambda_{\max} - \lambda$。因此，如果内积低于强规则界 $\lambda - (\lambda_{\max} - \lambda)$，变量就不能达到纳入模型所需要的水平，在这种情况下可以再次使用 KKT 条件。定义 $c_j(\lambda) = \boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}(\lambda))$，若全局规则或序列强规则成立，则有

$$\left|\frac{\mathrm{d}c_j(\lambda)}{\mathrm{d}\lambda}\right| \leqslant 1 \tag{5.80}$$

假定该导数存在①，则 λ 处的 KKT 条件为

$$c_j(\lambda) = \lambda s_j(\lambda)\ ,\quad j = 1, 2, \ldots, p \tag{5.81}$$

其中若 $\hat{\beta}_j(\lambda) \neq 0$，则 $s_j(\lambda) = \mathrm{sgn}(\hat{\beta}_j(\lambda))$；若 $\hat{\beta}_j(\lambda) = 0$，则 $s_j(\lambda) \in [-1, 1]$。根据链式求导法，可知

$$\frac{\mathrm{d}c_j(\lambda)}{\mathrm{d}\lambda} = s_j(\lambda) + \lambda \cdot \frac{\mathrm{d}s_j(\lambda)}{\mathrm{d}\lambda}$$

如果忽略掉上式的第二项，则有 $|\frac{\mathrm{d}c_j(\lambda)}{\mathrm{d}\lambda}| \leqslant 1$。现在当变量在 λ 取值的区间内有非零系数时，第二项会等于零，因为 $s_j(\lambda)$ 是一个常量（等于 ± 1）。除此以外，如果 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 是对角占优（diagonally dominant），则斜率条件（5.80）总是成立（Tibshirani et al. 2012），这个条件意味着特征几乎不相关。但一般情况下，斜率条件会因为 λ 的伸缩而达不到要求，从而导致强规则失效，即错误地丢掉特征。但这种情况很少见，当 $p \gg N$ 时，几乎不出现这种失效。

综上所述，根据经验发现，强规则，尤其是序列强规则（5.79）对丢掉特征具有很好的启发性。这些理论对 lasso、基于惩罚的逻辑斯蒂回归以及弹性网等都成立。

为了节省计算时间且不牺牲解的精确性，可按如下方式使用序列强规则：沿着递减 λ 值的细网络来求解。对于每个 λ 的值，采用筛选规则来得到特征子集，然后

① 这只是一种具有启发意义的讨论，因为后面的讨论会发现在 $\hat{\beta}_j(\lambda) = 0$ 处，$\mathrm{d}c_j(\lambda)/\mathrm{d}\lambda$，$\mathrm{d}s_j(\lambda)/\mathrm{d}\lambda$ 并不存在。

仅使用该子集来求解这个问题，再对所有特征检查看是否满足 KKT 条件（5.81）。如果满足，则求解成功；否则，会有不符合条件的特征加到活动集中，要再求解一次这个问题。原则上讲，没有特征违反条件才可停止求解。

不符合强规则条件的很少见（尤其是 $p \gg N$ 时），因此这种方法非常有效。Tibshirani et al.（2012）在 glmnet 中对坐标下降法实现了这些规则，并且在广义梯度法和 Nesterov 一阶方法中也实现了这些规则。加速因子给出的范围是 2~80，具体数值取决于设置。

最后来看看更一般的凸优化形式

$$\underset{\beta}{\text{minimize}}\left\{f(\beta)+\lambda\sum_{j=1}^{r}c_j\left\|\beta_j\right\|_{p_j}\right\} \tag{5.82}$$

其中 f 是可微凸函数。向量 $\beta=(\beta_1,\beta_2,\ldots,\beta_r)$。其中 β_j 是一个标量或向量。$\lambda \geqslant 0$，对于每个 $j=1,\ldots,r$，　$c_j \geqslant 0$，$p_j \geqslant 1$。对于给定的 $\lambda' > \lambda$，序列强规则为

$$\left\|\nabla_j f(\hat{\beta}(\lambda'))\right\|_{q_j} < c_j(2\lambda-\lambda') \tag{5.83}$$

其中 $\nabla_j f(\hat{\beta})=(\partial f(\hat{\beta})/\partial\beta_{j1},\ldots,\partial f(\hat{\beta})/\partial\beta_{jm})$，$1/p_j+1/q_j=1$（即 $\|\cdot\|_{p_j}$ 和 $\|\cdot\|_{q_j}$ 是对偶形式）。规则（5.83）可用于各种问题，包括逻辑斯蒂回归和其他广义线性模型，组 lasso 和图 lasso。

参考文献注释

下降算法的性质（包括基于适当步长选择规则来证明收敛性，如限制最小化或 Armijo 规则）是优化中的经典问题，更多详细内容参见 Bertsekas（1999）的第 1 章和第 2 章。关于拉格朗日方法和对偶的详细介绍可参考 Bertsekas（1999），以及 Boyd and Vandenberghe（2004）。Rockafellar（1996）对凸对偶和凸分析给出了更高级的处理。Nesterov（2007）Nesterov（2007）用广义梯度法（5.21）对组合目标函数进行了分析。Nesterov 的书（2004）对投影梯度法进行了相关分析。Lange（2004）以及 Hunter and Lange（2004）都介绍了优化–最小化过程（也称为辅助函数法）。

Gorski, Pfeuffer and Klamroth（2007）给出了双凸函数及优化它们的交替算法概述。El Ghaoui, Viallon and Rabbani（2010）介绍了筛选规则，如式（5.76）。基于该工作，我们得到了很类似的公式，并由些得到 5.10 节中介绍的强规则。最近，Wang, Lin, Gong, Wonka and Ye（2013）给出了更安全的规则，并提出了一个简单的序列化公式。Fu（1998）很早就在 lasso 中使用了坐标下降。

附录 A　lasso 的对偶

本附录会推导 lasso 原问题 (2.5) 的实用对偶形式，lasso 原问题的形式更简洁。

$$\text{lasso 原问题：} \underset{\beta\in\mathbb{R}^p}{\text{minimize}}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2+\lambda\|\beta\|_1 \tag{5.84}$$

引入残差向量 $\boldsymbol{r}=\boldsymbol{y}-\boldsymbol{X}\beta$，可重写原问题为

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\frac{1}{2}\|\boldsymbol{r}\|_2^2+\lambda\|\beta\|_1\,,\quad \boldsymbol{r}=\boldsymbol{y}-\boldsymbol{X}\beta \tag{5.85}$$

用 $\theta\in\mathbb{R}^N$ 来表示拉格朗日乘子向量，则该问题的拉格朗日函数为

$$L(\beta,\boldsymbol{r},\boldsymbol{\theta}):=\frac{1}{2}\|\boldsymbol{r}\|_2^2+\lambda\|\beta\|_1-\theta^{\mathrm{T}}(\boldsymbol{r}-\boldsymbol{y}+\boldsymbol{X}\beta) \tag{5.86}$$

最小化式（5.86）可得到关于 β 和 $\boldsymbol{r}$ 的对偶目标函数。将含有 β 的项单独列出来，则有

$$\min_{\beta\in\mathbb{R}^p}-\theta^{\mathrm{T}}\boldsymbol{X}\beta+\lambda\|\beta\|_1=\begin{cases}0, & \|\boldsymbol{X}^{\mathrm{T}}\theta\|_\infty\leqslant\lambda\\ -\infty, & \text{其他}\end{cases} \tag{5.87}$$

其中 $\|\boldsymbol{X}^{\mathrm{T}}\theta\|_\infty=\max\limits_j\left|\boldsymbol{x}_j^{\mathrm{T}}\theta\right|$。将含有 $\boldsymbol{r}$ 的项单独列出来，则有

$$\min_{\boldsymbol{r}}\frac{1}{2}\|\boldsymbol{r}\|_2^2-\theta^{\mathrm{T}}\boldsymbol{r}=-\frac{1}{2}\theta^{\mathrm{T}}\theta \tag{5.88}$$

其中 $\boldsymbol{r}=\theta$。将式（5.87）和式（5.88）代入拉格朗日函数式（5.86），则有

$$\text{lasso 对偶：} \underset{\boldsymbol{\theta}}{\text{maximize}}\frac{1}{2}\left\{\|\boldsymbol{y}\|_2^2-\|\boldsymbol{y}-\theta\|_2^2\right\}\,,\quad \|\boldsymbol{X}^{\mathrm{T}}\theta\|_\infty\leqslant\lambda \tag{5.89}$$

综上所述，lasso 的对偶形式可看成是将 $\boldsymbol{y}$ 投影到可行域集 $\mathcal{F}_\lambda=\left\{\theta\in\mathbb{R}^N\mid \|\boldsymbol{X}^{\mathrm{T}}\theta\|\leqslant\lambda\right\}$。$\mathcal{F}_\lambda$ 是 $2p$ 个由 $\left\{\left|\boldsymbol{x}_j^{\mathrm{T}}\theta\right|\leqslant\lambda\right\}_{j=1}^p$ 定义的半平面相交而得到的，它是 $\mathbb{R}^N$ 中的一个多面体。在 5.3.3 节中，这个解可由 $\theta^*=\mathbf{prox}_{I(\mathcal{F}_\lambda)}(\boldsymbol{y})$ 给出。图 5-13 给出了这个解几何解释的示意图。

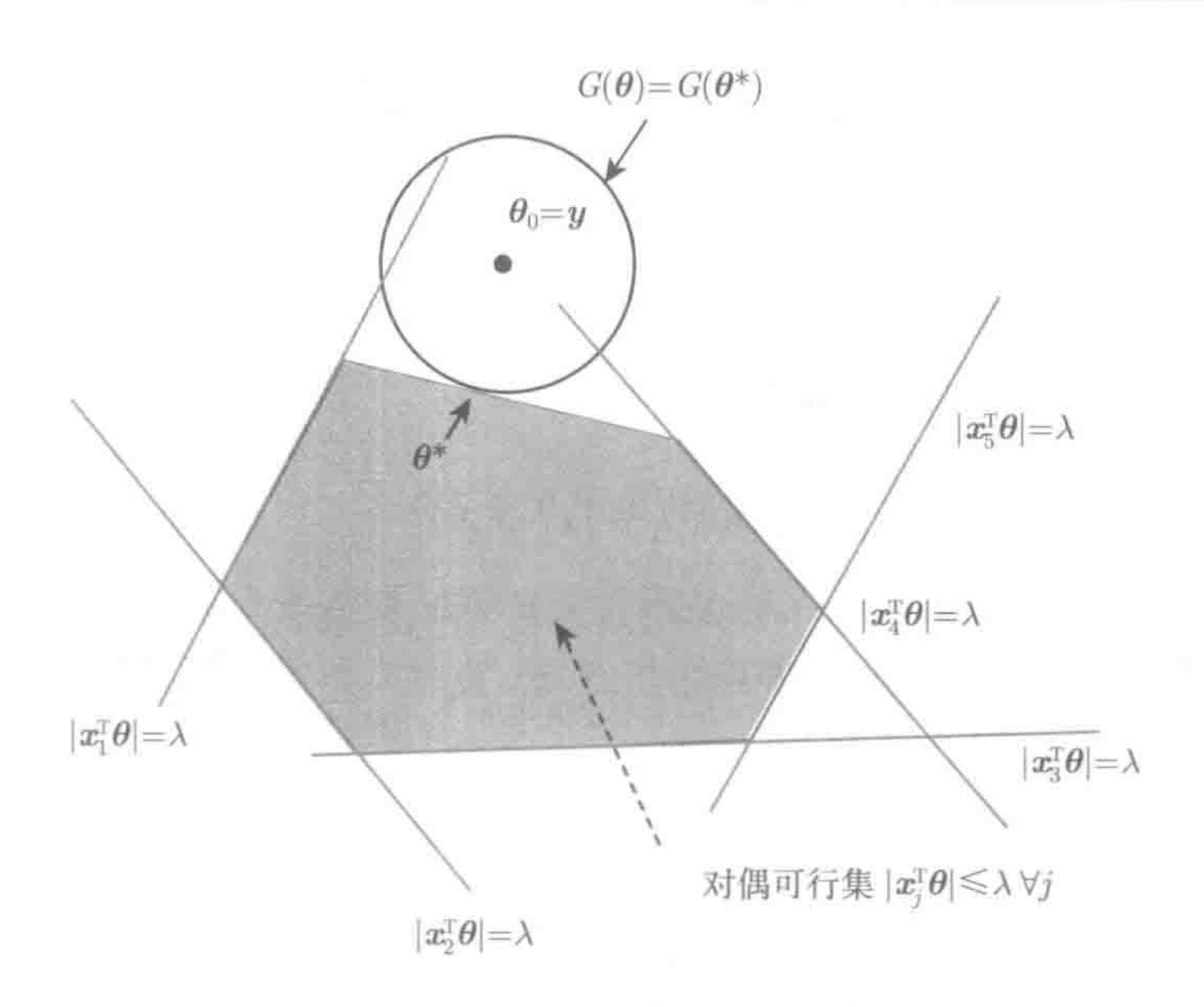


图 5-13　lasso 的拉格朗日对偶：$G(\theta)=\frac{1}{2}\left(\|\boldsymbol{y}\|_2^2-\|\boldsymbol{y}-\theta\|_2^2\right)$。阴影区域表示可行集 $\mathcal{F}_\lambda$。对偶的无约束解为 $\theta_0=\boldsymbol{y}$，会得到空残差。对偶解为 $\theta^*=\mathbf{prox}_{I(\mathcal{F}_\lambda)}(\boldsymbol{y})$，是 $\boldsymbol{y}$ 在凸集 $\mathcal{F}_\lambda$ 上的投影

附录 B　DPP 规则的推导

下面介绍序列 DPP 筛选规则（5.77）的推导过程证明来自 Wang, Lin, Gong, Wonka and Ye（2013）。先定义变量 $\boldsymbol{\phi}=\theta/\lambda$，然后替换掉 lasso 对偶中的 θ，则有

$$\underset{\boldsymbol{\theta}}{\text{maximize}}\,\frac{1}{2}\left\{\|\boldsymbol{y}\|_2^2-\lambda^2\,\|\boldsymbol{y}/\lambda-\boldsymbol{\phi}\|_2^2\right\},\qquad \left\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\phi}\right\|_\infty\leqslant 1 \tag{5.90}$$

定理 5.1：假定 $\lambda_{\max}\geqslant\lambda'>0$，lasso 对偶问题（5.90）的解为 $\hat{\boldsymbol{\phi}}(\lambda')$。设 $\lambda>0$ 且 $\lambda\neq\lambda'$，若不等式

$$\left|\boldsymbol{x}_j^{\mathrm{T}}\hat{\boldsymbol{\phi}}(\lambda')\right|<1-\|\boldsymbol{x}_j\|_2\,\|\boldsymbol{y}\|_2\left|\frac{1}{\lambda'}-\frac{1}{\lambda}\right| \tag{5.91}$$

成立，则有 $\hat{\beta}_j(\lambda)=0$。由于 $\hat{\boldsymbol{\phi}}(\lambda')=\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}(\lambda')\right)/\lambda'$，因此可得到式（5.77）。

证明：从 lasso 的稳定点条件可得

$$\left|\boldsymbol{x}_j^{\mathrm{T}}\hat{\boldsymbol{\phi}}(\lambda')\right|<1\Rightarrow\hat{\beta}_j(\lambda)=0 \tag{5.92}$$

从 lasso 的对偶式（5.90）可知，$\hat{\boldsymbol{\phi}}(\lambda)$ 是 $\boldsymbol{y}/\lambda$ 在可行集 $\mathcal{F}_\lambda$ 上的投影。由闭凸集上

的投影定理（Bertsekas 2003）可知，$\hat{\boldsymbol{\phi}}(\lambda)$ 连续且非可扩展。由此可以得出

$$\begin{aligned}\|\hat{\boldsymbol{\phi}}(\lambda)-\hat{\boldsymbol{\phi}}(\lambda')\|_2 &\leqslant \|\frac{\boldsymbol{y}}{\lambda}-\frac{\boldsymbol{y}}{\lambda'}\|_2\\ &=\|\boldsymbol{y}\|_2\left|\frac{1}{\lambda}-\frac{1}{\lambda'}\right|\end{aligned}\tag{5.93}$$

则有

$$\begin{aligned}\left|\boldsymbol{x}_j^{\mathrm{T}}\hat{\boldsymbol{\phi}}(\lambda)\right| &\leqslant \left|\boldsymbol{x}_j^{\mathrm{T}}\hat{\boldsymbol{\phi}}(\lambda)-\boldsymbol{x}_j^{\mathrm{T}}\hat{\boldsymbol{\phi}}(\lambda')\right|+\left|\boldsymbol{x}_j^{\mathrm{T}}\hat{\boldsymbol{\phi}}(\lambda')\right|\\ &<\left\|\boldsymbol{x}_j^{\mathrm{T}}\right\|_2\|\hat{\boldsymbol{\phi}}(\lambda)-\hat{\boldsymbol{\phi}}(\lambda')\|_2+1-\|\boldsymbol{x}_j\|_2\|\boldsymbol{y}\|_2\left|\frac{1}{\lambda'}-\frac{1}{\lambda}\right|\\ &\leqslant \|\boldsymbol{x}_j\|_2\|\boldsymbol{y}\|_2\left|\frac{1}{\lambda'}-\frac{1}{\lambda}\right|+1-\|\boldsymbol{x}_j\|_2\|\boldsymbol{y}\|_2\left|\frac{1}{\lambda'}-\frac{1}{\lambda}\right|=1\end{aligned}\tag{5.94}$$

习　　题

习题 5.1　已知最小化无约束的二次函数 $f(\beta)=\frac{1}{2}\beta^{\mathrm{T}}\boldsymbol{Q}\beta-\langle\beta,b\rangle$，其中 $\boldsymbol{Q}\succ 0$ 为对称正定矩阵，且 $b\in\mathbb{R}^p$。

(a) 求证：最优解 β^* 存在且唯一，写出其 $(\boldsymbol{Q},b)$ 的形式。

(b) 用具有固定步长 s 的梯度下降法求解这个问题的最优化解。

(c) 求证：存在某个常量 $c>0$（仅依赖于 $\boldsymbol{Q}$），对任意固定步长 $s\in(0,c)$，梯度下降收敛。

习题 5.2　最小化有约束的目标函数 $f(\beta)$，其约束条件为 $g_j(\beta)\leqslant 0(j=1,\dots,m)$，设 f^* 为最优值。定义拉格朗日函数为

$$L(\beta;\lambda)=f(\beta)+\sum_{j=1}^{m}\lambda_j g_j(\beta)\tag{5.95}$$

(a) 求证：

$$\sup_{\lambda\geqslant 0}L(\beta,\lambda)=\begin{cases}f(\beta), & g_j(\beta)\leqslant 0,\ j=1,\dots,m\\ +\infty, & \text{其他}\end{cases}$$

(b) 使用 (a) 中结果来证明 $f^*=\inf_\beta\sup_{\lambda\geqslant 0}L(\beta;\lambda)$

(c) f^* 与 $\sup_{\lambda\geqslant 0}L(\beta;\lambda)$ 之间有怎样的联系？

习题 5.3　设 $f:\mathbb{R}^p\to\mathbb{R}$ 是可微凸函数，对该函数有子空间约束 $\mathcal{C}=\{\beta\in\mathbb{R}^p|M\beta=c\}$，其中 $M\in\mathbb{R}^{m\times p}$ 是固定矩阵，$c\in\mathbb{R}^m$ 是固定向量。

(a) 假定 $\beta^*\in\mathcal{C}$ 满足一阶优化条件 (5.4)。求证必存在向量 $\lambda^*\in\mathbb{R}^m$ 使得

$$\nabla f(\beta^*)+M^{\mathrm{T}}\lambda^*=0\tag{5.96}$$

(b) 求证：若条件（5.96）对某个 $\lambda^* \in \mathbb{R}^m$ 成立，则一阶优化条件（5.4）一定成立。

习题 5.4　设与约束优化问题（5.5）相关的拉格朗日函数为 $L(\beta,\lambda) = f(\beta) + \sum_{j=1}^{m} \lambda_i g_j(\beta)$，设 f^* 为一个有限的优化值，并且存在两个向量 $\beta^* \in \mathbb{R}^p$ 和 $\lambda^* \in \mathbb{R}_+^m$，使得对于所有的 $\beta \in \mathbb{R}^p$ 和 $\lambda \in \mathbb{R}_+^m$，式

$$L(\beta^*,\lambda) \overset{\text{(i)}}{\leqslant} L(\beta^*,\lambda^*) \overset{\text{(ii)}}{\leqslant} L(\beta,\lambda^*) \tag{5.97}$$

成立求证：β^* 是该约束问题的最优解。

习题 5.5　**欧几里得范数的次梯度。**欧几里得范数也称 ℓ_2 范数，其定义为 $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$，用于组 lasso。

(a) 求证：对于 $\beta \neq 0$，范数 $g(\beta) := \|\beta\|_2$ 可微，其导数为 $\nabla g(\beta) = \dfrac{\beta}{\|\beta\|_2}$。

(b) 求证：对于 $\beta = 0$，任意的向量 $\hat{s} \in \mathbb{R}^p$ 且 $\|\hat{s}\|_2 \leqslant 1$ 是 g 在 0 处的次微分中的一个元素。

习题 5.6　求证式（5.40）中的

$$h(\beta_1,\ldots,\beta_p) = |\beta|^{\mathrm{T}} \boldsymbol{P} |\beta|$$

满足正则条件（5.39）。（可得出这样的结论：虽然式（5.40）中的函数不可分，但仍可用坐标下降法。）

习题 5.7　求证近点梯度法的迭代式（5.21）与式（5.19）相等。

习题 5.8　求证当 h 是给定的原子范数，可通过下面的过程来得到迭代式（5.26）。

(a) 计算输入矩阵 $\boldsymbol{Z}$ 的奇异值分解，即 $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$，其中 $\boldsymbol{D} = \operatorname{diag}\{\sigma_j(\boldsymbol{Z})\}$ 为奇异值构成的对角矩阵。

(b) 通过软阈值算子（5.25）来计算“收缩”的奇异值

$$r_j := S_{s\lambda}(\sigma_j(\boldsymbol{Z})),\ j = 1,\ldots,p$$

(c) 返回矩阵 $\hat{\boldsymbol{Z}} = \boldsymbol{U}\operatorname{diag}\{\gamma_1,\ldots,\gamma_p\}\boldsymbol{V}^{\mathrm{T}}$。

练习 5.9　对于回归问题，数据集中所有特征和输出向量的样本均值为 0，标准样本方差为 1。假设每个特征与输出向量相关的绝对值一样，即

$$\frac{1}{N}|\langle \boldsymbol{x}_j, \boldsymbol{y}\rangle| = \lambda\ ,\quad j = 1,\ldots,p$$

这里假设 $\hat{\beta}$ 是 $\boldsymbol{y}$ 在 $\boldsymbol{X}$ 上的唯一最小二乘系数向量。令 $\boldsymbol{u}(\alpha) = \alpha\boldsymbol{X}\hat{\beta}(\alpha \in [0,1])$ 为一个向量，可移动 α 到最小二乘所拟和的结果 $\boldsymbol{u}$。设 $\mathrm{RSS} = \|y - \boldsymbol{X}\beta\|_2^2$ 为整个最小二乘的残差的平方和。

(a) 求证：

$$\frac{1}{N}\left|\langle \boldsymbol{x}_j, \boldsymbol{y}-\boldsymbol{u}(\alpha)\rangle\right| = (1-\alpha)\lambda, \quad j=1,\dots,p$$

且当朝着 $\boldsymbol{u}$ 移动时，每个 $\boldsymbol{x}_j$ 与残差的相关系数大小相同。

(b) 求证这些相关系数都等于

$$\lambda(\alpha) = \frac{(1-\alpha)}{\sqrt{(1-\alpha)^2 + \dfrac{\alpha(2-\alpha)}{N}\cdot \mathrm{RSS}}}\cdot \lambda$$

并且会单调递减至 0。

(c) 使用这些结果来证明 5.6 节的 LAR 算法与这些相关系数有联系，并且 LAR 的系数单调递减。

练习 5.10 对于算法 5.1 中的步骤 3c，令 $c_l = \langle \boldsymbol{x}_l, \boldsymbol{r}_{k-1}\rangle$，$a_l = \langle \boldsymbol{x}_l, \boldsymbol{X}_A\delta, \rangle, \ell \notin A$，另外还定义

$$\alpha_\ell = \min\nolimits_+\left\{\frac{\lambda_{k-1}-c_\ell}{1-a_\ell}, \frac{\lambda_{k-1}+c_\ell}{1+a_\ell}\right\}$$

其中 $\min_+$ 只取为正的元素。求证：第 k 步的变量的索引 $j=\arg\min_{l\notin\mathcal{A}}\alpha_\ell$ 的值为 $\lambda_k = \lambda_{k-1}-\alpha_j$。

练习 5.11 强规则

(a) 求证：如果斜率条件（5.80）成立，则全局强规则（5.78）和序列强规则（5.79）有效。

(b) 假设有正交条件 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}=\boldsymbol{I}$，求证：斜率条件（5.80）总是成立。

(c) 针对 lasso 设计一个模拟实验，研究 DPP 和强规则的精度。要求 $(N,\ p)=(100, 20)$，$(N,\ p)=(100, 100)$，$(N,\ p)=(100, 1000)$。

练习 5.12 用 ADMM 求解一致优化。有数据集 $\{x_i, y_i\}_{i=1}^N$，我们需要最小化目标函数 $L(\boldsymbol{X}\beta-\boldsymbol{y})$，它可分解成 N 项相加，每一项对应一个样本。此处自然可以将数据集分成 B 个块，在数据的第 b 个块上所对应的目标函数可表示为 $L_b(\boldsymbol{X}_b\beta_b-\boldsymbol{y}_b)$，其中，$\boldsymbol{X}_b$ 和 $\boldsymbol{y}_b$ 分别对应 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的第 b 个块。该问题可以重写为

$$\underset{\beta\in\mathbb{R}^p}{\operatorname{minimize}}\left\{\sum_{b=1}^{B} L_b\left(\boldsymbol{X}_b\beta_b-\boldsymbol{y}_b\right)+r(\theta)\right\},\ \beta_b=\theta,\ b=1,\dots,B \tag{5.98}$$

(a) 求证对这个问题所采用的 ADMM 算法应形如：

$$\beta_b^{t+1} \leftarrow \arg\min_{\beta_b}\left(L_b\left(\boldsymbol{X}_b\beta_b-\boldsymbol{y}_b\right)+(\rho/2)\left\|\beta_b-\theta^t+\mu_b^t\right\|_2^2\right) \tag{5.99a}$$

$$\theta^{t+1} \leftarrow \arg\min_{\theta} \left(r(\boldsymbol{z}) + (N\rho/2) \left\| \theta - \bar{\beta}^{t+1} - \bar{\mu}^{t} \right\|_2^2 \right) \tag{5.99b}$$

$$\mu_b^{t+1} \leftarrow \mu_b^t + \left(\beta_b^{t+1} - \theta^{t+1} \right) \tag{5.99c}$$

其中 $\bar{\mu}^k$和$\bar{\beta}^{k+1}$ 表示整个块的平均值。这可以称为一致优化。

(b) lasso 采用的正则项为 $r(\theta) = \lambda \|\theta\|_1$，求证针对 lasso 的 ADMM 算法的迭代公式形如

$$\beta_b^{t+1} \leftarrow \left(\boldsymbol{X}_b^T \boldsymbol{X}_b + \rho \boldsymbol{I} \right)^{-1} \left(\boldsymbol{X}_b \boldsymbol{y}_b + \rho \left(\theta^t - \mu_b^t \right) \right) \tag{5.100a}$$

$$\theta^{t+1} \leftarrow \mathcal{S}_{\lambda/(\rho N)} \left(\bar{\beta}^{t+1} + \bar{\mu}^t \right) \tag{5.100b}$$

$$\mu_b^{t+1} \leftarrow \mu_b^t + \left(\beta_b^{t+1} - \theta^{t+1} \right) \tag{5.100c}$$

(c) 用软件实现迭代式（5.100），并在相关的数据上证明结果。

习题 5.13 (a) 针对问题（5.72）推导交替迭代凸最小化，并证明它有一个 power 迭代形式 [式（5.73）和式（5.74）]。

(b) 求证在初始向量 $\boldsymbol{v}_0$ 不与矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 最大特征对应的特征向量正交的情况下，ACS 会收敛到该特征向量。

第6章　统 计 推 断

ℓ_1 正则化方法能够让特征选择和参数拟合同时进行，这是一个非常吸引人的特性。通常用交叉验证来选择模型（即对预测或泛化误差进行估计），然后在一个独立测试集上做进一步的验证。

有时人们感兴趣的是计算模型中变量的统计强度，如传统模型中的 p-value。估计方法的自适应特性使得这个问题（在概念上和分析上）变得困难。本章会介绍一些推断问题的实用方法首先讨论两种“传统”的方法，即贝叶斯方法和自助法（bootstrap），然后介绍解决这些问题新方法。

6.1　贝叶斯 lasso

贝叶斯方法把参数视为随机变量，这些随机变量具有先验分布，表示这些变量在人们心目中的取值。这里采用 Park and Casella（2008）的方法，其模型为

$$\boldsymbol{y}\,|\beta,\lambda,\sigma \sim N(\boldsymbol{X}\beta,\sigma^2\boldsymbol{I}_{N\times N}) \tag{6.1a}$$

$$\beta|\lambda,\sigma \sim \prod_{j=1}^{p}\frac{\lambda}{2\sigma}\mathrm{e}^{-\frac{\lambda}{\sigma}|\beta_j|} \tag{6.1b}$$

式（6.1b）是独立同分布的拉普拉斯分布。在这个模型下，很容易证明 $\beta|\boldsymbol{y},\lambda,\sigma$ 的负对数后验概率密度为

$$\frac{1}{2\sigma^2}\left\|\boldsymbol{y}-\boldsymbol{X}\beta\right\|_2^2+\frac{\lambda}{\sigma}\left\|\beta\right\|_1 \tag{6.2}$$

这里去掉了独立于 β 的一个常数。因此，对于任意给定的 σ 和 λ 值，后验模式和 lasso 估计（正则化参数为 $\sigma\lambda$）同时发生。Park and Casella（2008）因为技术原因将 σ^2 包含于先验概率里面 [见式（6.1b）]。这里给出的模型并没有常数，$\boldsymbol{X}$ 的每一列都进行了均值中心化，如同在 $\boldsymbol{y}$ 中一样。① 后验分布所能提供的不仅仅是点估计，还有整个联合分布。

图 6-1 中的红线是贝叶斯 lasso 中用到的双指数先验分布，运用在“糖尿病数据”中的变量 β_7 上。数据集包含 442 个样本，响应变量是一个定量的测量值，是在基准后一年的疾病级数（disease progression）。数据集有十个基本变量：年龄、性别、体重系数、平均血压和 6 组血清测量值，再加上二次项，共有 64 个特征。先验

① 在模型中这不是一个真正的约束，等价于在 β_0 上假设了一个不当且平的先验，这基本上没有什么意义。

分布在 0 处有一个尖峰值，这使人们相信有一些参数会为零。在给定参数时，为观测数据假定一个概率分布（似然函数），然后通过调节观测数据来更新先验，从而产生参数的后验分布。图 6-1 中的直方图描述了糖尿病数据中 β_7 的后验分布。先验分布有一个方差参数，从中可以看出人们确信 0 是一个特殊值。虽然 95%的后验置信区间包含了零，但后验模型有一点偏离零值。精确的贝叶斯计算很难获得，因此人们会使用最简单的贝叶斯模型。幸运的是，现代计算机能够用马尔可夫链蒙特卡洛（Markov Chain Monte Carlo，MCMC）方法从重点参数的后验分布中进行采样实现。图 6-2 左图在后验分布 $\beta|\lambda$ 中进行 MCMC 采样；图中展示了每 100 个 λ 值上 10000 个后验采样值的中值。这里 σ^2 是可变的（有 $\pi(\sigma^2)\sim\frac{1}{\sigma^2}$）。事实上这里显示了中位数，这与右图（lasso）有细微差别，右图显示了固定 $\sigma\lambda$ 值的后验模式。完整的贝叶斯模型会指定 λ 的先验分布。在这里，Gamma 分布是共轭的，因此适合用于 MCMC 采样。这就是很多时候人们选择贝叶斯方法的原因。完整的后验分布包括 λ 和 β，所以会自动进行模型选择。此外，β 的后验置信区间考虑了在 λ 上的后验可变性。图 6-3 汇集了从糖尿病数据的后验采样中所得到的 10000 个 MCMC 样本。尽管模型中有 9 个非零系数，但是后验分布认为其中只有 5~8 个不为零。

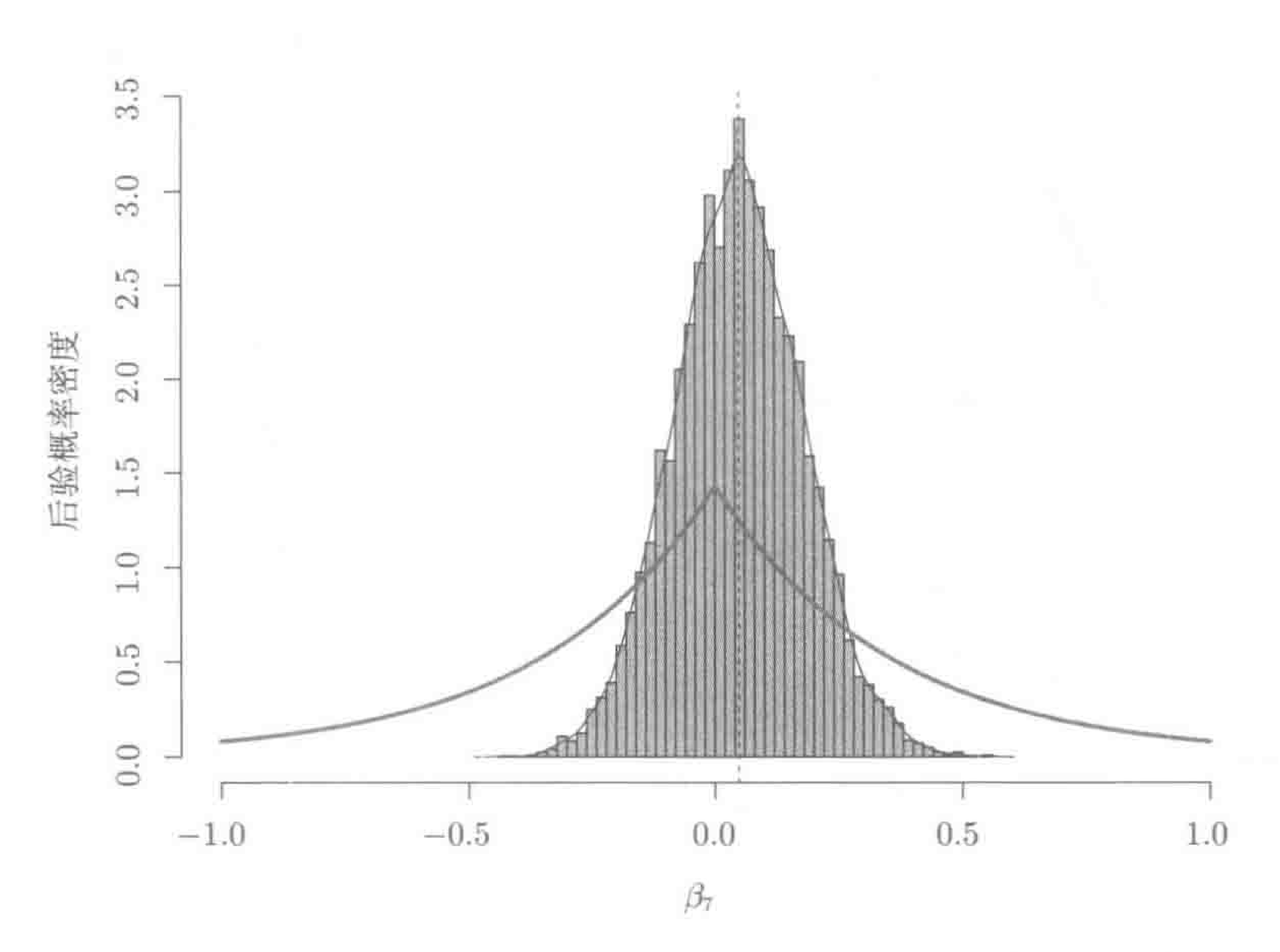


图 6-1　以糖尿病为例，当 λ 固定时，第七个变量的先验和后验分布。图中的先验分布是双指数分布，密度函数为 $\exp(-0.0065|\beta_7|)$。假设先验比率为 0.0065

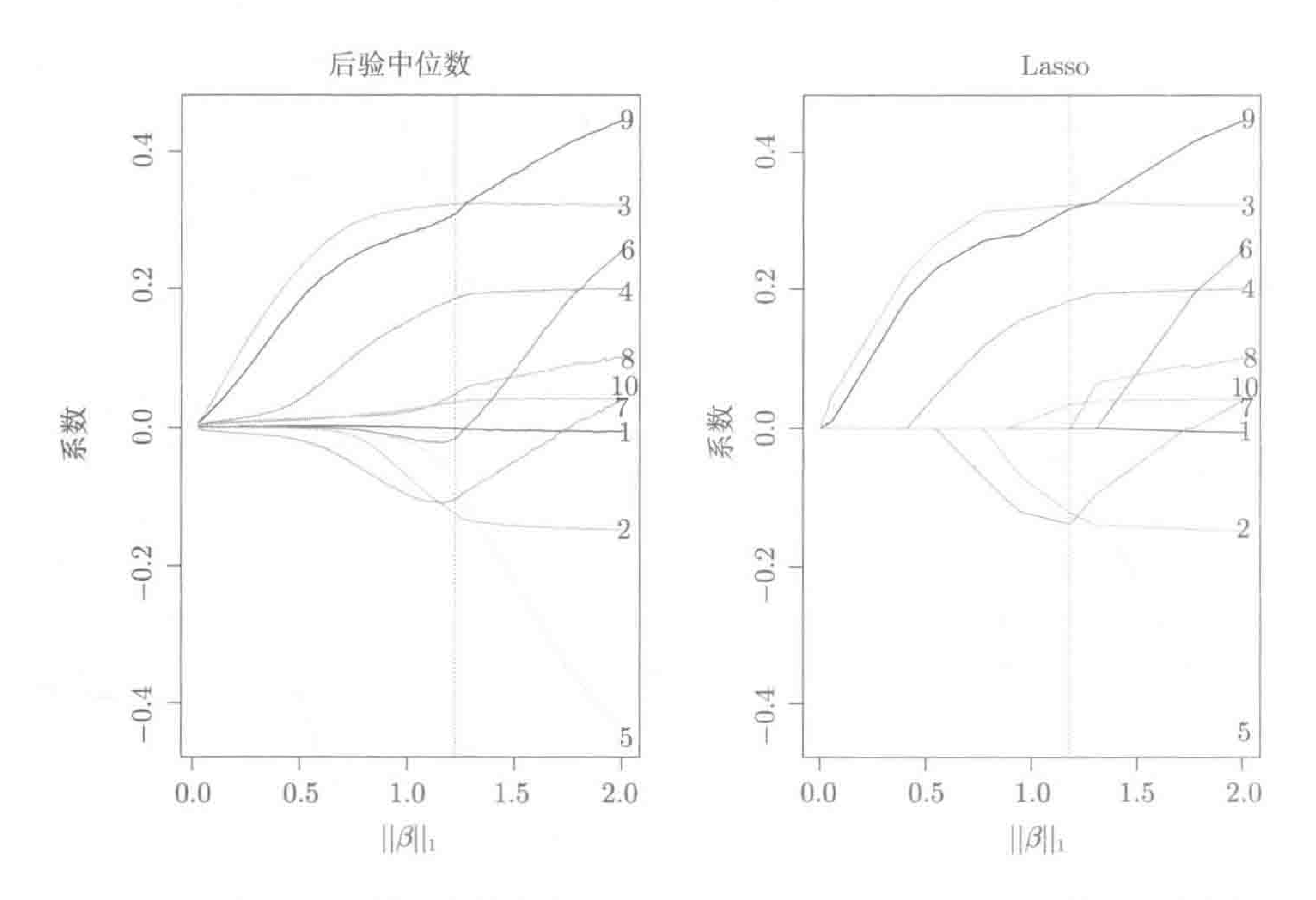


图 6-2　贝叶斯 lasso 在糖尿病数据上的应用。左图为 MCMC 方法得到的后验中位数（基于 λ）。右图为 lasso 方法。左图中的垂直线位于 $||\beta||_1$ 的后验中位数处（来自一个无条件模型），右图中的垂直线是通过 N 折交叉验证得到的

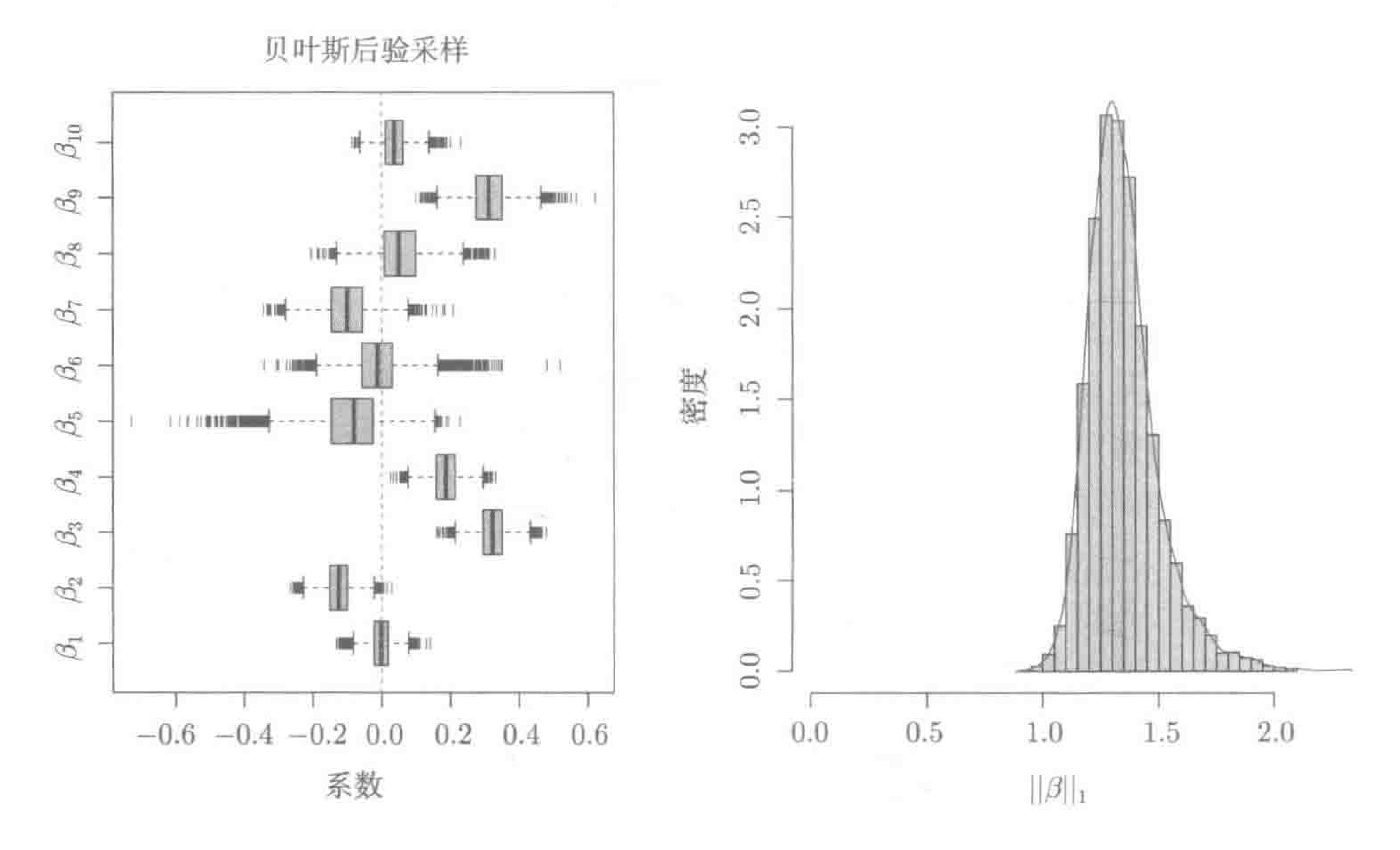


图 6-3　糖尿病数据中 β_j 和 $||\beta||_1$ 的后验分布。这里给出了 10 000 个 MCMC 样本，舍弃了前 1000 个“老化”（burn-in）样本

确定贝叶斯模型是一项技术挑战，要进行很多选择，其中涉及 λ 和 σ^2 的先验分布，而先验分布本身又需要确定超参数。这里的例子是用 R 语言包 monomvn（Gramacy 2011）中的函数 blasso 来拟和的，一般采用默认参数设置。对这个 442×10 问题，程序在一台 2.3GHz MacBook Pro 上运行完花了 5 秒。但是，贝叶斯计算的可扩展性并不好，6.2 节的实验会说明计算的代价大致为 $\mathcal{O}(p^2)$ 。

6.2 自 助 法

自助法是一种常用的非参数方法，用来评价复杂估计量的统计特性（Efron 1979, Efron and Tibshirani 1993）。为了引出自助法，假设通过以下步骤得到一个 lasso 问题的估计 $\widehat{\beta}(\widehat{\lambda}_{CV})$。

(1) 通过一个稠密的网格值 $\Lambda=\{\lambda_\ell\}_{\ell=1}^L$ 对数据（$\boldsymbol{X}, \boldsymbol{y}$）拟合出一条 lasso 路径。

(2) 随机将训练样本分为 10 组。

(3) 留下第 k 组，在剩余的 9/10 组中用同样的网格 Λ 拟合 lasso 路径。

(4) 在留下的那一组上，对每个 $\lambda \in \Lambda$ 计算均方预测误差。

(5) 平均这些误差，得到一条网格 Λ 上的预测误差曲线。

(6) 找到曲线取得最小值时的 $\widehat{\lambda}_{CV}$ 值，然后到第 (1) 步中找到在此 λ 值下拟合得到的系数向量。

如何评估采样分布 $\widehat{\beta}(\widehat{\lambda}_{CV})$? 这里的重点是将随机估计 $\widehat{\beta}(\widehat{\lambda}_{CV})$ 的分布作为 N 个独立同分布样本 $\{(x_i, y_i)\}_{i=1}^N$ 的函数。非参数自助法是这种采样分布的近似方法。为了做到近似，它通过 N 个样本定义的经验 CDF$\widehat{F}_N$ 来近似随机对 (X, Y) 的累积分布函数 F，然后我们从 $\widehat{F}_N$ 中抽出 N 个样本，相当于从给定数据集中有放回地抽取 N 个样本。图 6-4 左图是基于这种方法的 1000 次自助实现的 $\widehat{\beta}^*(\widehat{\lambda}_{CV})$ 箱线图，在每次自助采样中重复步骤 (1)~(6)。[①]该图与图 6-3 中的贝叶斯结果之间有合理的对应关系。右图为自助分布中各个变量为零的次数比。尽管贝叶斯后验方法得到的结果十分接近于零，但是没有一个完全为零。[blasso 函数有一个参数，允许通过可逆跳跃（reversible jump）的 MCMC 方法来进行变量选择，但是这里并没有使用。] 与右图相似，Meinshausen and Buhlmann（2010）画出了自助重采样下 lasso 的**稳定**图；作为 λ 的函数，他们给出了在自助法下系数路径中变量的非零次数比。

图 6-5 为自助法下的交叉验证曲线和它们的最小值。不出所料，自助法下的最小值有很宽的范围，因为原始的 CV 曲线在很宽的域上都是平的。有趣的是，自助法的标准差带与左图中原始的 CV 拟合计算得到的标准差带有很强的对应关系。

① 从技术上讲要设置样本权重为 $w_i^* = k/N$，其中 $k = 0, 1, 2, \ldots$，如此实行自助法。在交叉验证中，单元重复原先的 N 个样本，并会携带它们的权重 w_i^*。

图 6-6 为自助法下系数的成对图。从该图可以看到相关变量之间是如何权衡的，这里同时考虑了值及其为零的倾向。

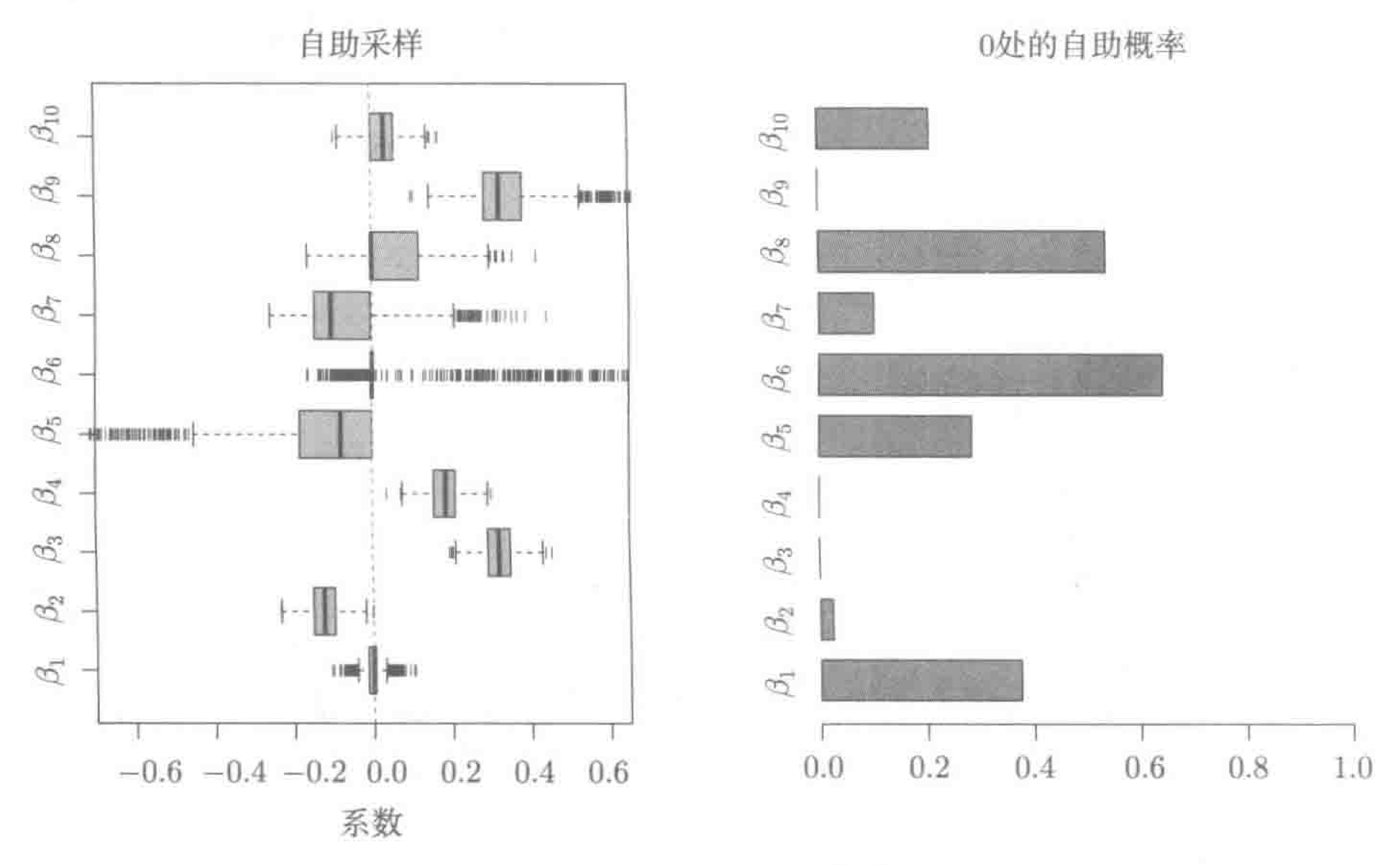


图 6-4　左图为非参数自助法下 1000 组观测样本的 $\widehat{\beta}^*(\widehat{\lambda}_{CV})$ 箱线图，对应经验 CDF $\widehat{F}_N$ 中的重采样。与图 6-3 中对应的贝叶斯后验分布相比，这种情况的对应要密切一些。右图为自助分布中各个参数为 0 的次数比

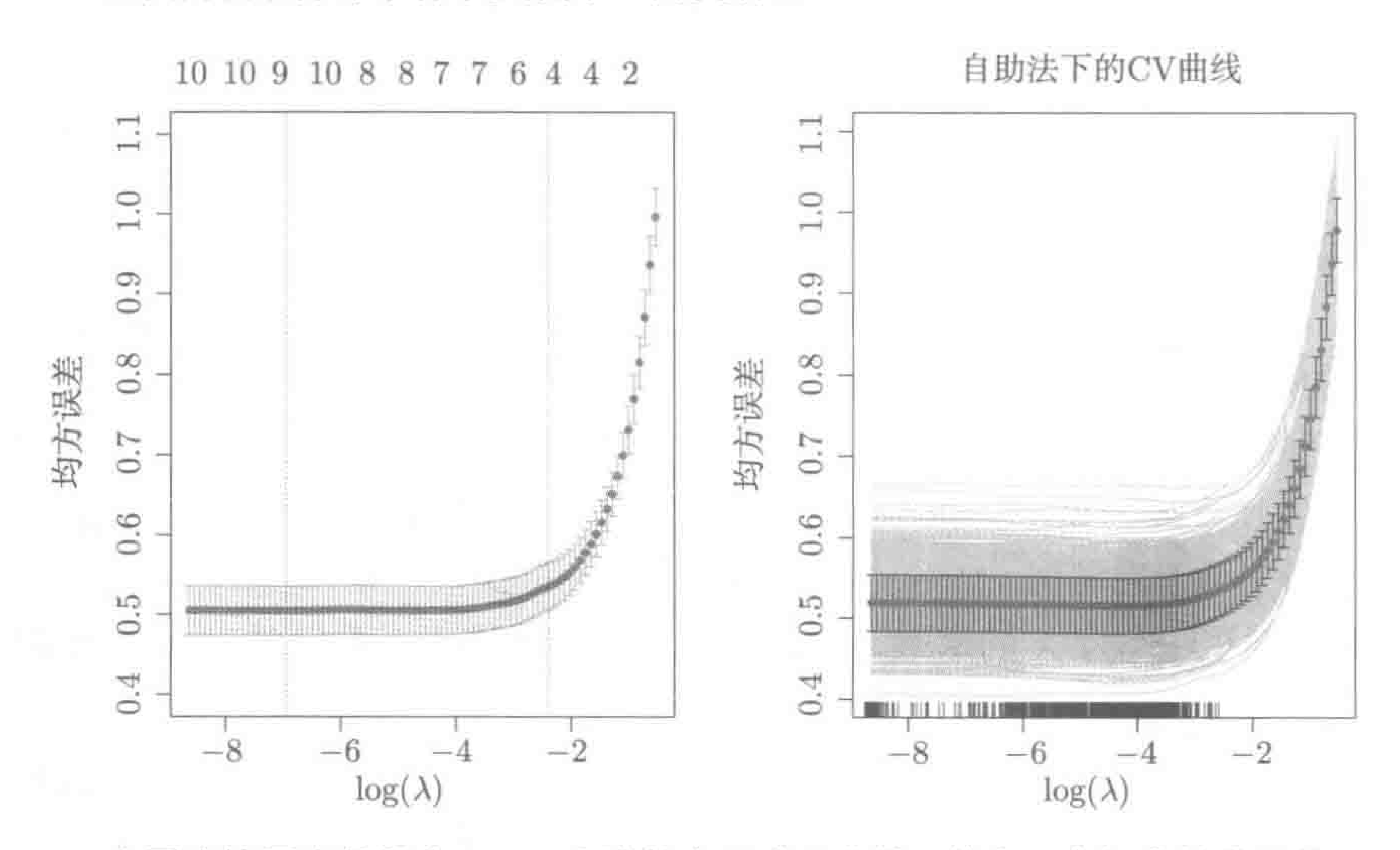


图 6-5　左图为糖尿病数据上 lasso 方法的交叉验证曲线，其中一个标准差带是从 10 个观测值计算得到的。左边的垂直线对应 λ 的极小值。右边的曲线对应一个标准差带规则，在 CV 误差中最大的 λ 是最小值的一个标准差。右图为 1000 组自助法下的 CV 曲线，红色是平均线，蓝线是一个标准差带。底下的毯状线对应着取极小值的地方

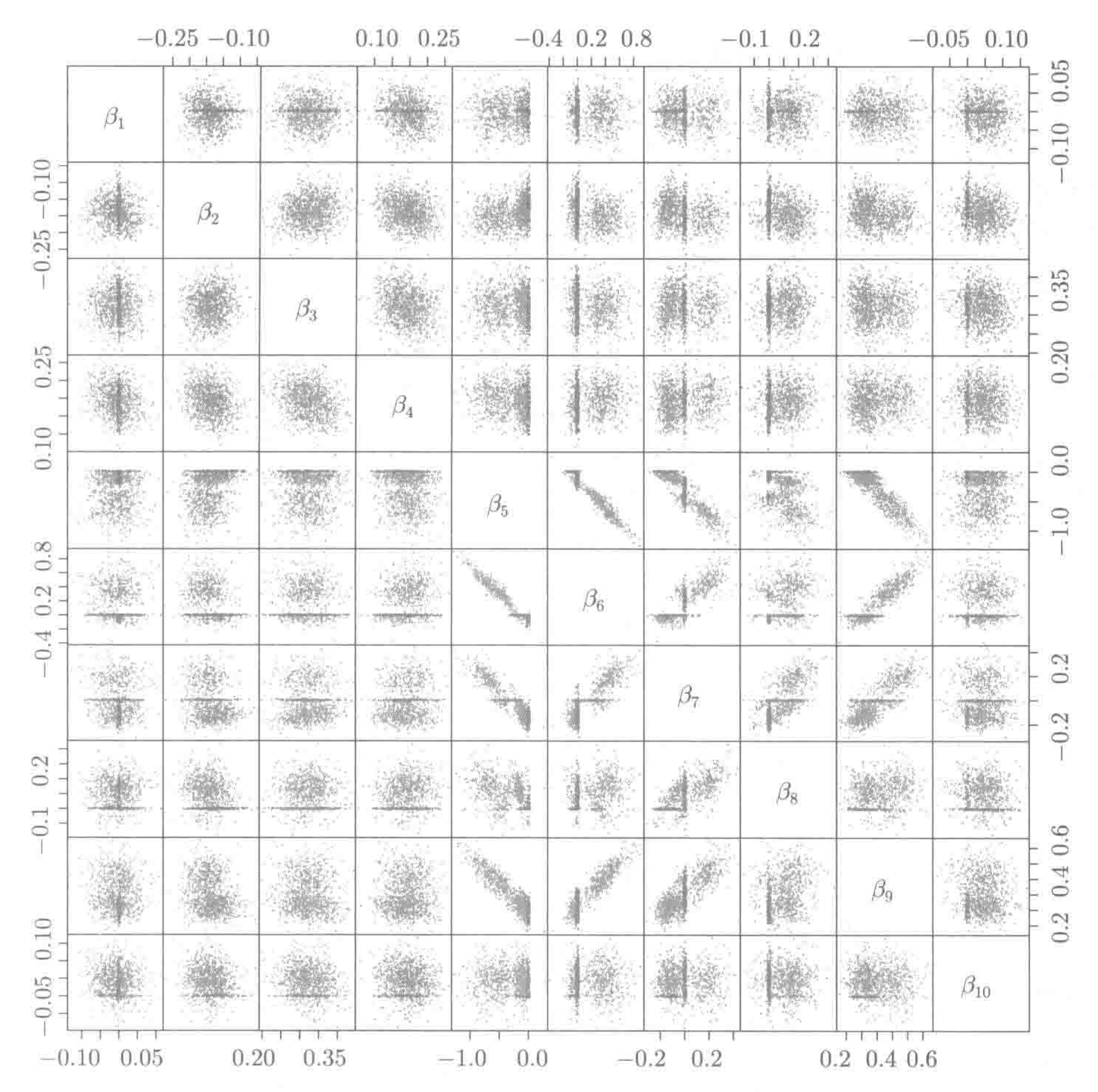


图 6-6 自助法系数 $\widehat{\beta}^*(\widehat{\lambda}_{CV})$ 的成对图。图中红点表示至少在一个坐标轴上值为零。样本 x_5 和 x_6 有很强的相关性（相关系数为 0.9）；从图中可以看到它们的系数是负相关，很多地方都为 0

表 6-1 在问题规模为 $N=400$ 时，对于不同数目的预测子进行了时间比较（以秒为单位）。这里产生了 1000 组自助样本。对贝叶斯 lasso，这里生成了 2000 组后验样本，其中将前 1000 个样本视为老化样本而丢弃。这种比较依赖于实现细节，因此 p 的相对增长是带有信息的。贝叶斯 lasso 可能对小样本来说更快，但其计算复杂度为 $\mathcal{O}(p^2)$。而自助法的计算复杂度接近于 $\mathcal{O}(p)$，因为它利用了 lasso 的稀疏性和凸性。

上面的算法用了**非参数自助法**，通过经验分布函数 $\widehat{F}_N$ 来估计未知总体 F，$\widehat{F}_N$ 是 F 的非参数最大似然估计。从 $\widehat{F}_N$ 中采样对应从数据中有放回采样。相反，**参数自助法**是从 F 的参数估计中采样，或者说，它对应密度函数 f。本例给定 $\boldsymbol{X}$，从

完全最小二乘拟合或者带参数 λ 的 lasso 拟合中得到的 $\widehat{\beta}$ 和 $\widehat{\sigma}^2$ 的估计。接下来从高斯模型 (6.1a) 对 $\boldsymbol{y}$ 值采样，用 β 和 σ^2 替换 $\widehat{\beta}$ 和 $\widehat{\sigma}^2$。

采用完全最小二乘来估计 $\widehat{\beta}$ 和 $\widehat{\sigma}^2$，图 6-7 展示了参数自助法得到的结果。这个结果与非参数自助法的结果以及贝叶斯 lasso 的结果相似。通常人们可能期望与非参数自助法相比，参数自助法的结果要更加接近贝叶斯 lasso，因为参数自助法和贝叶斯 lasso 都对数据分布用到了假设的参数形式：$\boldsymbol{y}\,|\beta,\lambda,\sigma \sim N(\boldsymbol{X}\beta,\sigma^2\boldsymbol{I}_{N\times N})$。另外需要注意，当 $p \gg N$ 时，无法采用完全最小二乘拟合来估计 $\widehat{\beta}$ 和 $\widehat{\sigma}^2$，这需要对每个 λ 值生成不同的数据集。这将极大地降低计算效率。

表 6-1　不同维度下，贝叶斯 lasso 和自助法 lasso 的运行时间。样本数为 $N = 400$

p	贝叶斯 lasso	自助法 lasso
10	3.3 秒	163.8 秒
50	184.8 秒	374.6 秒
100	28.6 分钟	14.7 分钟
200	4.5 小时	18.1 分钟

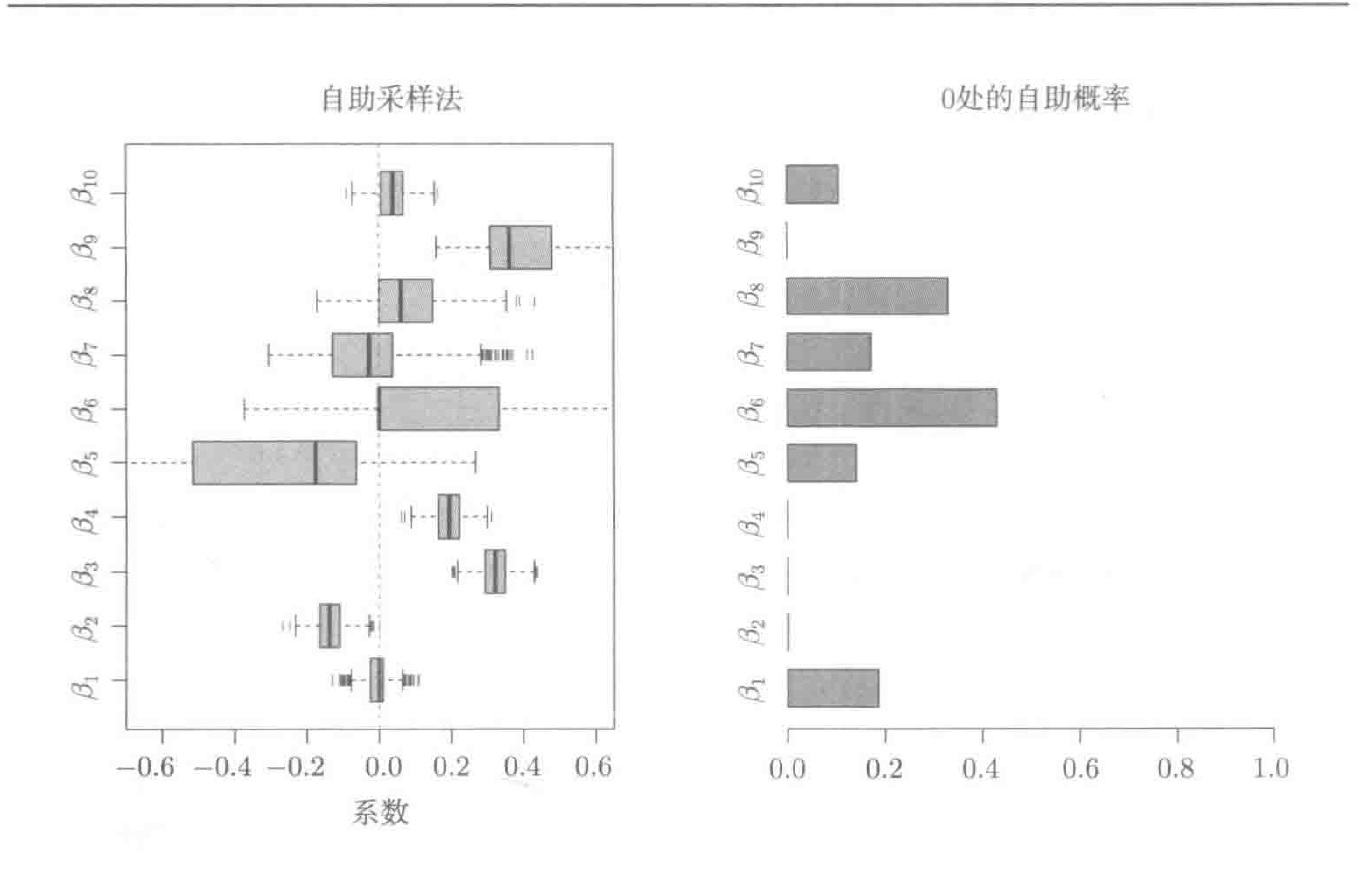


图 6-7　左图为参数自助法下 1000 组观测样本的 $\widehat{\beta}^*(\widehat{\lambda}_{CV})$ 箱线图。与图 6-3 中对应的贝叶斯后验分布相比，我们再次看到对应密切。右图为自助分布中各个系数为零的次数比

本节在高斯线性回归问题上比较了贝叶斯方法（在本文写作时有相关软件）和自助法。当考虑 GLM 和其他模型时，贝叶斯方法的计算复杂度增加。自助法可以

用在很多场合。一般情况下，贝叶斯 lasso 和 lasso/自助法会得到相似的结果。非参数自助法下得到的值的直方图，可看成是多项式模型在无先验信息下的一种后验贝叶斯估计（Rubin 1981，Efron 1982）。

哪一种方法更好呢？贝叶斯和自助法都是用来评价 lasso 估计的方法。贝叶斯方法原则性更强，但是与非参数自助法相比，过多地倚靠参数假设。自助法可以更好地扩展到大型数据集上。Efron（2011）给出了关于贝叶斯方法和自助法关系的深入讨论。

6.3 lasso 法的后选择推断

本节将介绍关于选择后进行推断的新观点，这些选择是通过自适应方法，比如 lasso 和向前逐步回归得到的。6.3.1 节讨论的第一种方法是基础，这种方法的一系列的推广和发展将在 6.3.2 节中介绍。

6.3.1 协方差检验

本小节介绍的方法可用于指定预测子的 p 值，这种方法相继被 lasso 等方法所采用，其基础是 LAR 算法及其分段求解 lasso 解的路径的思想（5.6 节）。

假设在通常的线性回归求解中，有一个输出向量 $\boldsymbol{y} \in \mathbb{R}^N$ 和预测子变量矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，两者之间的关系为

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N\times N}) \tag{6.3}$$

其中 $\beta \in \mathbb{R}^p$ 是待估计的未知系数。

为了理解方差检验的目的，首先来考虑向前逐步回归。这个算法一次只引入一个预测子，在每一步中选择能最大程度降低残差平方和的预测子。假设模型含有 k 个预测子，RSS_k 表示模型的残差平方和，用残差平方和的变化可得到检验统计量

$$R_k = \frac{1}{\sigma^2}(\text{RSS}_{k-1} - \text{RSS}_k) \tag{6.4}$$

（这里假设 σ 已知），将其与 χ_1^2 分布比较。

图 6-8a 为向前逐步回归中的 R_1 分位数（模型中第一个预测子的卡方统计量）与一个 χ_1^2 变量的分位数进行的比较，这些都是在完全空模型下（β=0）进行的。观察到的分位数要比 χ_1^2 分布中的大得多。例如在 5%上的检验，使用 3.84 作为 χ_1^2 截止值，将会有大约 39%的实际类型 I 的误差。

其原因十分清楚：卡方检验假设要比较的模型是预先指定的，并非基于数据进行了选择。但是，向前逐步回归的目的是在所有的变量中选择最好的预测子，所以它的训练误差要比期望低很多。

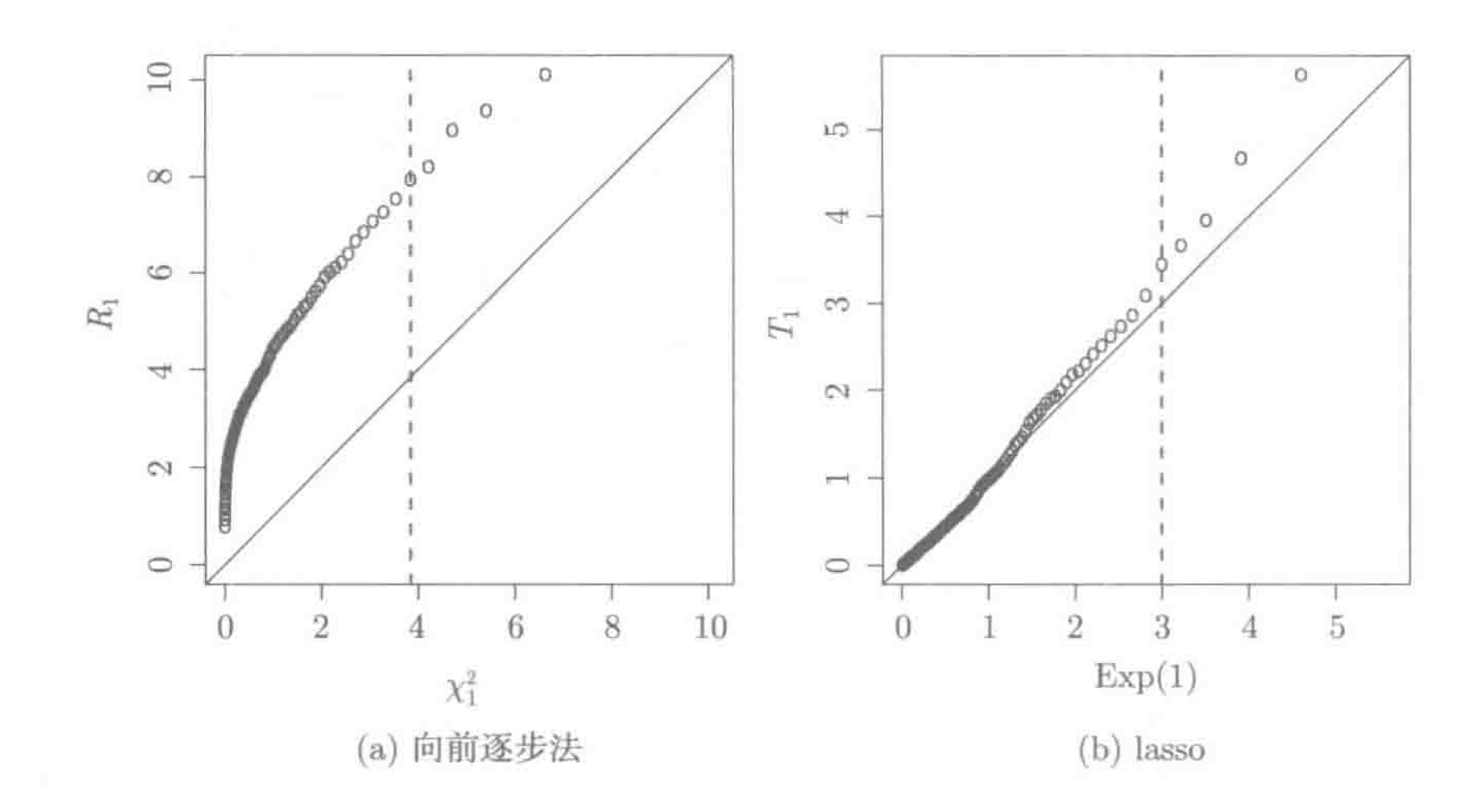


图 6-8　一个仿真例子，有 $N=100$ 个样本，$p=10$ 个正交预测子，β=0。(a) 为 1000 次模拟下的式（6.4）标准卡方统计量 R_1 对 χ_1^2 分布的分位数–分位数图。在向前逐步回归中，加入第一个预测子后，通过残差平方和的下降来度量 R_1。虚垂直线为 χ_1^2 分布的 95%分位数线。(b) 在式 (6.5) 中第一个加入 lasso 路径预测子的方差检验统计量 T_1 与及其渐近零分布 Exp(1) 的分位数–分位数图。协方差检验说明了 lasso 建模的自适应特性，而在向前逐步回归中卡方检验并不合适在自适应选择模型中应用

向前逐步回归法很难得到合适的 p 值，这是其拟合的自适应性所决定的。对于第一步及全局空假设检验，可采用置换分布。后继步骤如何正确进行置换步骤并不清楚。可以采用样本分割：将数据分为两半，对一半计算一系列模型，对另一半评估模型的显著性。这会大幅损失计算精度，除非样本数量十分大。

令人惊讶的是，lasso 可通过简单的检验来解释自适应性。用 $\lambda_1>\lambda_2\ldots>\lambda_K$ 表示 LAR 算法（算法 5.1）中返回的节点值，这些便是正则化参数 λ 的值，在那里起作用变量的数据集会发生变化。假设要检验 λ_k 处 LAR 算法选择的预测子的显著性。令 $\mathcal{A}_{k-1}$ 表示该预测子加入之前的活动集（系数非零的预测子），这一步最后的估计为 $\widehat{\beta}(\lambda_{k+1})$。重新拟合 lasso，让 $\lambda=\lambda_{k+1}$，但只采用 $\mathcal{A}_{k-1}$ 中的变量。由此将产生估计量 $\widehat{\beta}_{\mathcal{A}_{k-1}}(\lambda_{k+1})$。**方差检验估计量**的定义为

$$T_k=\frac{1}{\sigma^2}\cdot\left(\left\langle\boldsymbol{y},\boldsymbol{X}\hat{\beta}(\lambda_{k+1})\right\rangle-\left\langle\boldsymbol{y},\boldsymbol{X}\hat{\beta}_{A_{k-1}}(\lambda_{k+1})\right\rangle\right)\tag{6.5}$$

这种统计量会度量计算结果和拟合模型之间的协方差对刚进入模型的预测子有多大贡献。例如，基于这种度量下，在区间 $(\lambda_k,\lambda_{k+1})$ 上的增加量是多少。有趣的是，对向前逐步回归法，对应的方差统计量等于式（6.4）的 R_k，而 lasso 的情况并非如此（见习题 6.2）。

根据零假设（即所有的 $k-1$ 显著变量都在模型中）和一般情况下的模型矩阵 $\boldsymbol{X}$，对于在下一步中加入模型中的预测子，随着 $N,p\to\infty$，有

$$T_k \xrightarrow{d} \text{Exp}(1) \tag{6.6}$$

图 6-8b 为 T_1 对 Exp(1) 的分位数–分位数图。当 σ^2 未知时，用全模型来估计 $\widehat{\sigma}^2=\dfrac{1}{N-p}\text{RSS}_p$。然后将这一项代入式（6.5）中，则指数检验变为 $F_{2,N-p}$ 检验。

表 6-2 为向前逐步回归和 LAR/lasso 在糖尿病数据上得到的结果。每种情形只展示了前十步。可以看到，向前逐步回归在 0.05 水平下有 8 项，而协方差检验中仅有 4 项。但向前逐步回归的 p 值偏低，所以并不怎么可信，这在前面也提到过。习题 6.3 会讨论一种方法，将一组连续的 p 值结合起来用于控制所选择预测子集的错误发现率（False Discovery Rate，FDR）。在 FDR 为 5%采用协方差检验时，会得到包含前 4 个预测子的模型。为了便于比较，交叉检验估计用于预测的最优模型大小为 7~14 个预测子。

为什么向前逐步统计量 R_1 的均值比 1 大得多，而 T_1 的均值近似于 1 呢？原因在于**收缩**：lasso 在每一步中选择出目前可用的最佳预测子，但是并没有通过最小二乘法拟合。它采用收缩的系数估计，这种收缩能补偿由于选择带来的膨胀。这个检验与 2.5 节中的 lasso 和 LAR 的自由度结果很类似。有 k 个非零系数的 lasso 希望自由度为 k，LAR 在解径的每一段 $(\lambda_k,\lambda_{k+1})$ 会用一个自由度。协方差检验的均值等于 1，这就是每一步的自由度。在某种意义上对自适应拟合而言，Exp(1) 分布与 χ_1^2 分布近似。

表 6-2 向前逐步回归和 LAR/lasso 法应用在第 2 章介绍的糖尿病数据上的结果。每种情形只显示了前面 10 步的结果。p 值分别基于式（6.4）、式（6.5）和式（6.11）。值 <0.01 则标为 0

向前逐步回归			LAR/lasso		
步骤	项	p 值	项	p 值	
				协方差	间隔
1	bmi	0	bmi	0	0
2	ltg	0	ltg	0	0
3	map	0	map	0	0.01
4	age:sex	0	hdl	0.02	0.02
5	bmi:map	0	bmi:map	0.27	0.26
6	hdl	0	age:sex	0.72	0.67
7	sex	0	glu^2	0.48	0.13
8	glu^2	0.02	bmi^2	0.97	0.86
9	age^2	0.11	age:map	0.88	0.27
10	tc:tch	0.21	age:glu	0.95	0.44

协方差检验式（6.5）的指数极限分布需要数据矩阵 $\boldsymbol{X}$ 有某些条件，即信号变量（有非零系数）与噪声变量并非十分相关。这些条件与 lasso 中（详见第 11 章）恢复所需要的条件相似。下一小节会讨论一种适用面更广的方案，即**间距检验**（spacing test），它的空分布对有限的 N 和 p 完全成立，并且对任意的 $\boldsymbol{X}$ 成立。

6.3.2 后选择推断的更广方案

这里将讨论后选择推断的一个更广方案：在高斯分布下可以计算出精确的 p 值和置信区间。该方案可以处理任意选择法，只要该选择法可以用一组 $\boldsymbol{y}$ 上的线性不等式来表示。换句话说，选择事件可以写成 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$的形式，其中 $\boldsymbol{A}$ 为矩阵，b 为向量。特别是，这可以用在 LAR 算法的一系列步骤中，给出协方差检验的精确（有限个样本）形式。同样，也可以用在向前逐步回归算法和有固定的正则化参数 λ 的 lasso 算法中。

为什么这种选择事件能够写成 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$的形式呢？从向前逐步回归的角度看，这一点十分明显。在这种情况下，可以将让 $b=0$。在第一步中，向前逐步回归选择的预测子与 $\boldsymbol{y}$ 的绝对内积最大（见图 6-10）。这可由一个有 $2(p-1)$ 行的矩阵 $\boldsymbol{A}$ 来表示，矩阵中每一行计算得到两内积之差，一个为正方向，一个为反方向。同样，在下一步中增加 $2(p-2)$ 行对比选择的预测子和其余的 $p-2$ 个预测子之间的内积，后面以此类推。

在给定 λ 值后，lasso 解能被一组变量 $\mathcal{A}$ 的活动集及其系数符号表示。此外，事实证明导致这种特定组合的选择事件对某些 $\boldsymbol{A}$ 和 b 可以写成 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$的形式。换言之，集合 $\{\boldsymbol{y}|\boldsymbol{A}\boldsymbol{y} \leqslant b\}$对应着输出向量 $\boldsymbol{y}$ 的值，向量 $\boldsymbol{y}$ 将产生相同的活动变量集及符号（$\boldsymbol{X}$ 固定时）（见 Lee, Sun, Sun and Taylor 2013 和习题 6.10）。对于 LAR 算法，第 k 步之后同样如此。

现在假设 $\boldsymbol{y} \sim N\left(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_{N\times N}\right)$，并且需要在事件 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$下作出推断。具体而言，要推断 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$，其中 $\boldsymbol{\eta}$ 可能依赖于选择事件。当 lasso、LAR 或者向前逐步回归已经选择出集合时，可以对选择出的变量进行推断。例如，人们可能对 $\boldsymbol{y}$ 在 $\boldsymbol{X}_{\mathcal{A}}$ 上的（普通）回归系数（即 $\widehat{\theta} = (\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}})^{-1}\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{y}$）有兴趣。这些对应着总体参数 $\theta = (\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}})^{-1}\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\mu}$，即 $\boldsymbol{\mu}$ 在 $\boldsymbol{X}_{\mathcal{A}}$ 上投射的系数。所以，这里 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$ 对应这些系数中的一个，因此 $\boldsymbol{\eta}$ 是 $\boldsymbol{X}_{\mathcal{A}}(\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}})^{-1}$ 中的一列。后面会继续研究这个例子。

Lee, Sun, Sun and Taylor(2013) 以及 Taylor, Lockhart, Tibshirani$_2$ and Tibshirani(2014) 证明

$$\{\boldsymbol{A}\boldsymbol{y} \leqslant b\} = \left\{\mathcal{V}^-(\boldsymbol{y}) \leqslant \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y} \leqslant \mathcal{V}^+(\boldsymbol{y}), \mathcal{V}^0(\boldsymbol{y}) \geqslant 0\right\} \tag{6.7}$$

而且，$\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$ 和 $\left(\mathcal{V}^-(\boldsymbol{y}), \mathcal{V}^+(\boldsymbol{y}), \mathcal{V}^0(\boldsymbol{y})\right)$ 在统计上是相互独立的。图 6-9 给出了这个结果的几何视图，即**多面体引理**（polyhedral lemma）。式（6.7）中右边的三个值可

以通过计算

$$
\begin{aligned}
\alpha &= \frac{\boldsymbol{A}\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|_2^2} \\
\mathcal{V}^-(y) &= \max_{j:\alpha_j<0} \frac{b_j - (\boldsymbol{A}\boldsymbol{y})_j + \alpha_j \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}}{\alpha_j} \\
\mathcal{V}^+(y) &= \min_{j:\alpha_j>0} \frac{b_j - (\boldsymbol{A}\boldsymbol{y})_j + \alpha_j \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}}{\alpha_j} \\
\mathcal{V}^0(y) &= \min_{j:\alpha_j=0} (b_j - (\boldsymbol{A}\boldsymbol{y})_j)
\end{aligned}
\tag{6.8}
$$

得到（见习题 6.7）。因此，选择事件 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$等价于事件 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$ 落进某一个范围，这个范围与 $\boldsymbol{A}$ 和 b 有关。这种等价性和独立性意味着关于 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$ 的条件性推断可以通过 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$ 的截尾分布来得到，这是一个截尾正态分布。

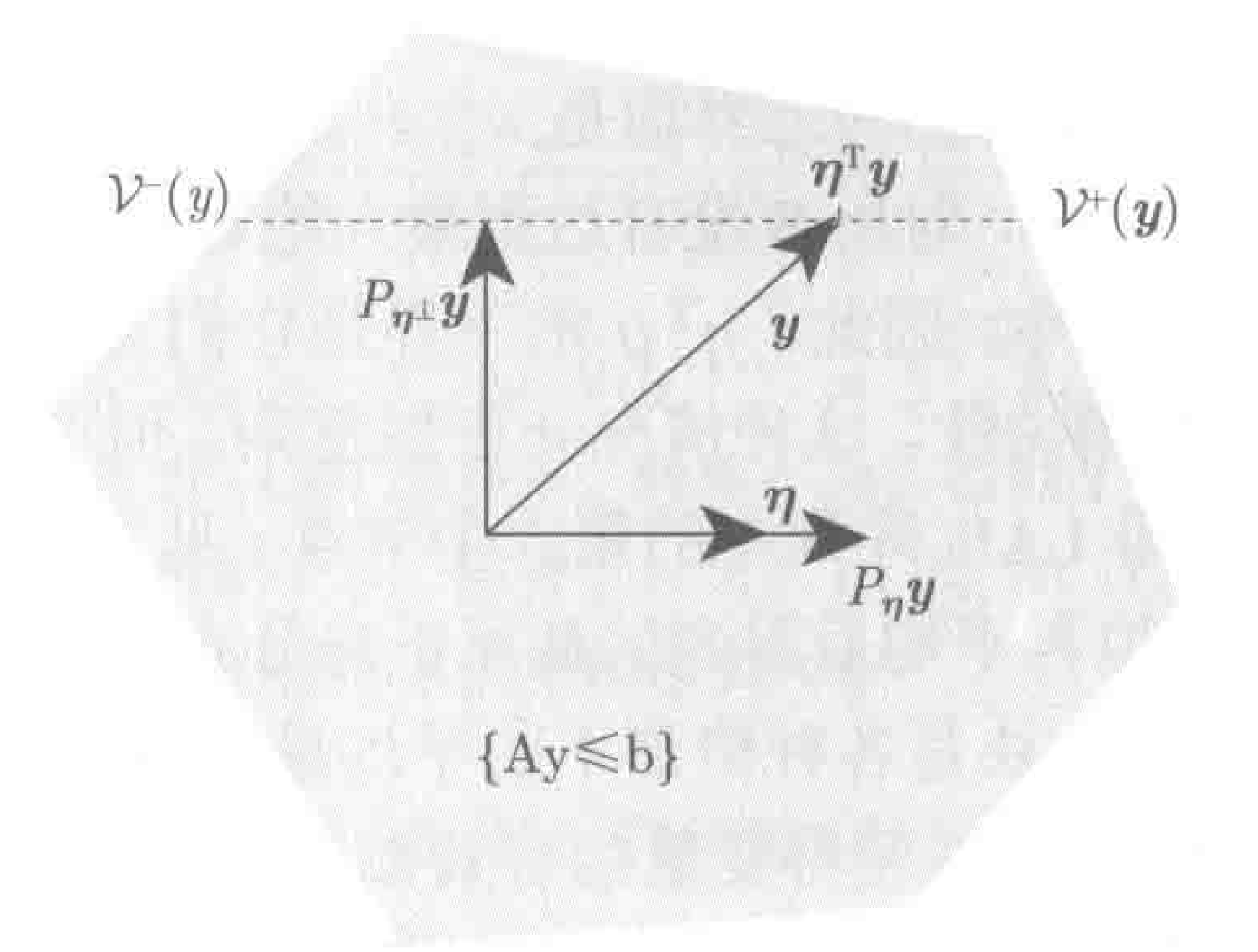


图 6-9　对于 $N=2$，$\|\boldsymbol{\eta}\|_2=1$ 的情形，多面体引理（6.7）的示意图。阴影区域是选择事件 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$。我们分解 $\boldsymbol{y}$ 为两项之和，即 $P_{\boldsymbol{\eta}}\boldsymbol{y}$ 在 $\boldsymbol{\eta}$（及坐标轴 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$）上的投射和在 $(N-1)$ 维子空间上的投射，该子空间正交于 $\boldsymbol{\eta}: \boldsymbol{y} = P_{\boldsymbol{\eta}}\boldsymbol{y} + P_{\boldsymbol{\eta}}^{\perp}\boldsymbol{y}$。基于条件 $P_{\boldsymbol{\eta}}^{\perp}\boldsymbol{y}$，可以看到事件 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$等价于事件 $\{\mathcal{V}^-(\boldsymbol{y}) \leqslant \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y} \leqslant \mathcal{V}^+(\boldsymbol{y})\}$。而 $\mathcal{V}^+(\boldsymbol{y})$ 和 $\mathcal{V}^-(\boldsymbol{y})$ 独立于 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$，因为前者只是 $P_{\boldsymbol{\eta}}^{\perp}\boldsymbol{y}$ 的函数，不受 $\boldsymbol{y}$ 影响

利用这一特性，可定义 $[c,d]$ 上的截尾正态分布的累积分布函数为（Cumulative Distribution Function，CDF）

$$
F_{\boldsymbol{\mu},\sigma^2}^{c,d}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((c-\mu)/\sigma)}{\Phi((d-\mu)/\sigma) - \Phi((c-\mu)/\sigma)} \tag{6.9}
$$

其中 Φ 是标准高斯分布的 CDF。现在，随机变量的 CDF（可得到该随机变量的概率）是一个均匀分布，因此有

$$
F_{\boldsymbol{\eta}^T\boldsymbol{\mu},\sigma^2\|\boldsymbol{\eta}\|_2^2}^{\mathcal{V}^-,\mathcal{V}^+}(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}) \,|\{\boldsymbol{A}\boldsymbol{y} \leqslant \boldsymbol{b}\} \sim U(0,1) \tag{6.10}
$$

这个结果可用于对任意线性函数 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$ 进行条件推断。例如，可以计算检验 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}=0$ 的 p 值。也可以通过对检验做如下转换来对 $\theta = \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$ 建立一个 $1-\alpha$ 的选择

区间。设 $P(\theta)=F_{\theta,\sigma^2\|\boldsymbol{\eta}\|_2^2}^{\nu^-,\nu^+}(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y})\,|\,\{\boldsymbol{A}\boldsymbol{y}\leqslant b\}$。区间的下界对应 θ 的最大值，即 $1-P(\theta)\leqslant\alpha/2$，区间的上界对应 θ 的最小值，即 $P(\theta)\leqslant\alpha/2$。

例 6.1 为了详细解释这些结果，这里举例说明。从模型 $Y=\sum_{j=1}^{p}X_j\beta_j+Z$ 中采样得到 $N=60$ 个样本，其中 $X_1,X_2,\ldots,X_p$，$Z\sim N(0,1)$，归一化样本，使其均值为 0，并使它们的 ℓ_2 范数为 1 。对于全局空假设下全部 β_j=0，预测子 j_1 和 $\boldsymbol{y}$ 之间有最大的绝对内积。这是第一个进入 LAR 和 lasso 解路径的变量。需要推断出 λ_1，这是在全局空假设下 LAR 中最大节点的值。因此 $\boldsymbol{\eta}=\boldsymbol{x}_{j_1}$ 和 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$ 是得到的内积（内积为正）。注意，由于归一化，$\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}=\boldsymbol{x}_{j_1}^{\mathrm{T}}\boldsymbol{y}$ 也是 $\boldsymbol{y}$ 在选择 $\boldsymbol{x}_{j_1}$ 上的简单最小二乘系数，这里也针对 $\boldsymbol{y}$ 与 $\boldsymbol{x}_{j_1}$ 上的简单回归进行了总体系数的（条件）推断。我们选择五种情形，预测子数目分别为 $p\in\{2, 5, 10, 20, 50\}$。有两种预测子关联模式纳入考虑：不关联和成对关联，相应的相关系数为 0.5。图 6-10 为图 6-9 在这两种情况所对应的版本。所有情况的上界均是 $\nu^+=\infty$，下界 ν^- 依赖于各个模拟中的 $\boldsymbol{y}$。在正交情况下（左图），对所有没有达到最大绝对内积的预测子 k，以 $P_{\boldsymbol{\eta}^{\perp}}\boldsymbol{y}$ 为条件变成以 $|\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{y}|$ 为条件。因此，$\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$ 的下界是它们当中第二高的。右图为非正交情况，X_j 之间具有相关性。这种情况会更加复杂，但可用一个简单公式来推导 $\nu^-(\boldsymbol{y})$，结果为 λ_2，即 LAR 序列中的第二个节点（见习题 6.11）。图 6-11 为对式（6.10）进行 100 次模拟再平均后得到的截尾正态分布密度函数。这里绘出了 ν^- 平均值密度。底部的彩色方块是各种情形下的平均最大内积 $\lambda_1=\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}$。在图 6-11 的下半部，随着 p 变大，关联性使得变量的有效数目更小，因此最大值也就变得更小。后面会继续讨论该例。

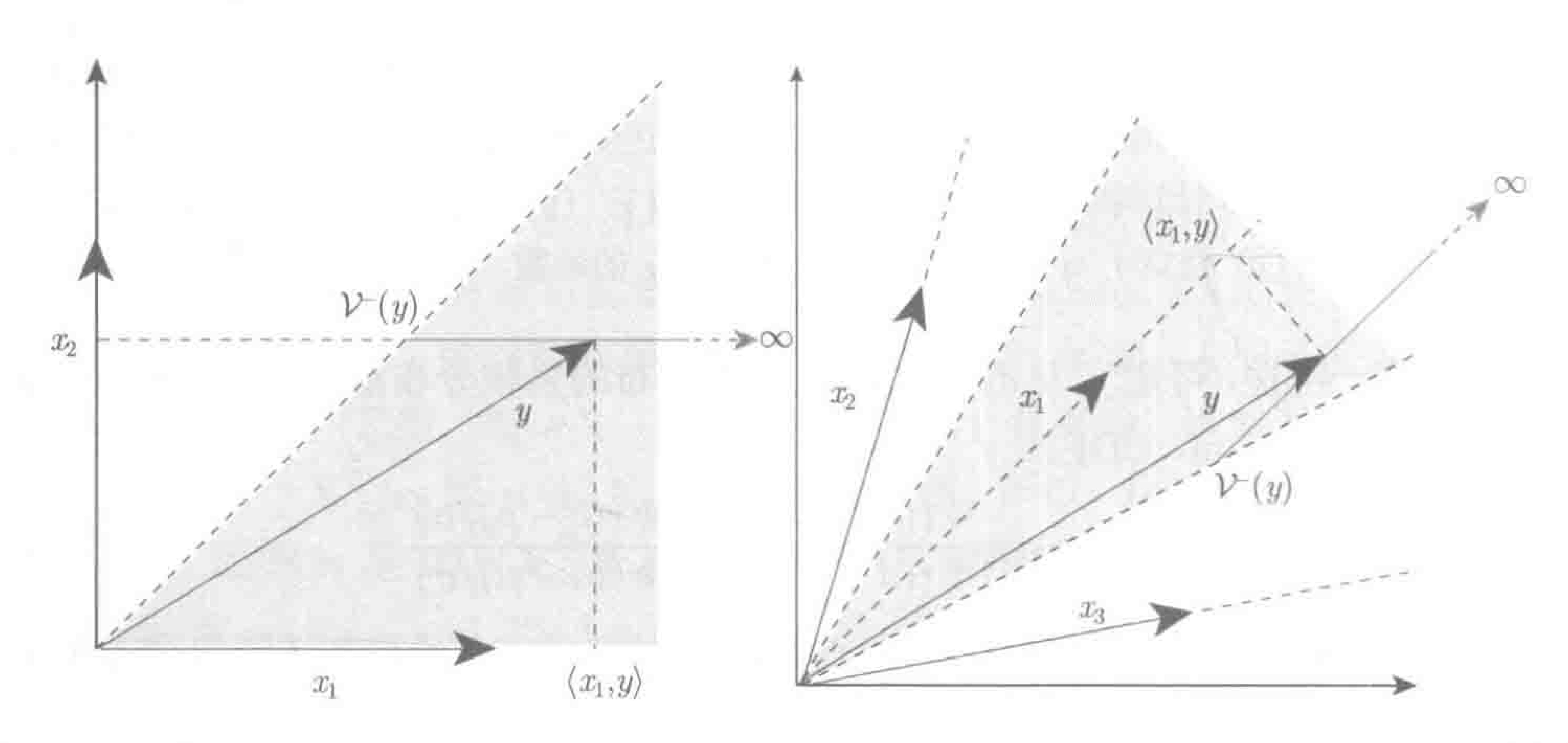


图 6-10 在 $\lambda_1=\langle\boldsymbol{x}_1,\boldsymbol{y}\rangle$ 时，例 6.1 的选择区域。左图所示为两个正交预测子，右图所示为三个相关预测子。红线表示集合 $P_{\boldsymbol{\eta}^{\perp}}\boldsymbol{y}+t\boldsymbol{\eta}$ 在选择区域内的部分。左图中 $\nu^-(\boldsymbol{y})=\langle\boldsymbol{x}_2,\boldsymbol{y}\rangle$，而右图中 $\nu^-(\boldsymbol{y})=\lambda_2$（见彩插）

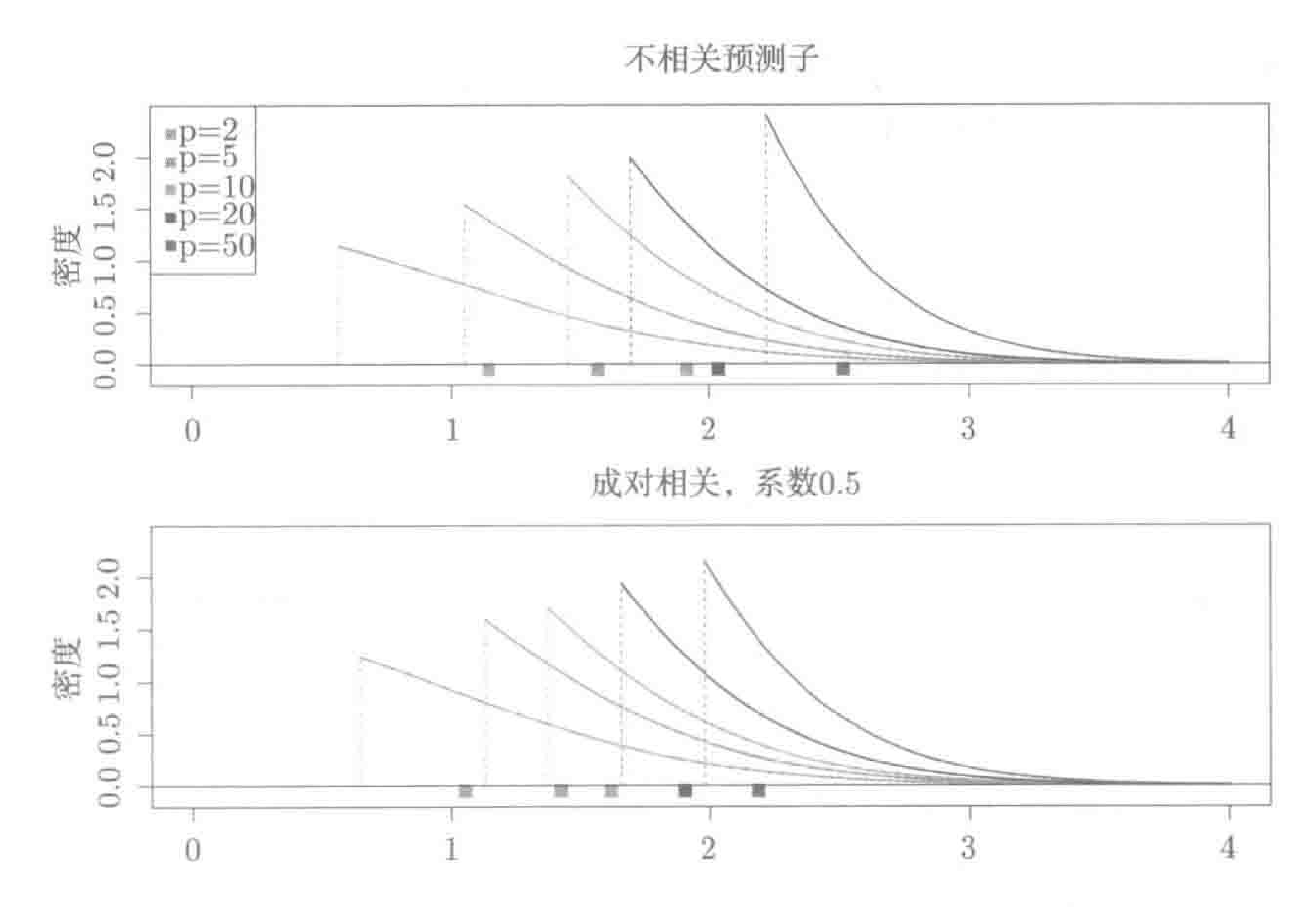


图 6-11 模拟：从模型 $X_1, X_2, \ldots, X_\mu \sim N(0,1) Y = \sum_j X_{ij}\beta_j + Z$，$Z \sim N(0,1)$，所有 β_j=0 中模拟得出 $N = 60$ 个样本，两种不同的预测子相关。选择事件挑选出与 $\boldsymbol{y}$ 有最大绝对内积的预测子 j_1。图为 $p = 2, 5, 10, 20, 50$ 时式（6.10）左边的截尾密度。底部的彩色方块为每种情形下的平均最大内积

这种通用机制 [见式（6.10）] 允许人们推断任意线性函数 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$。例如，在 LAR 算法的给定步骤中或者在 λ 处计算 lasso 解时，可推断任意参数 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$。根据设置不同，$\boldsymbol{A}$ 和 b 的形式会不同，除此之处，其结构是一样的。接下来举例说明。

1. **固定 λ，推断** lasso

下面介绍在固定 λ 值时 lasso 的解。构建 $\boldsymbol{A}$ 和 b 可应用式（6.10）的结果，即事件 $\{\boldsymbol{A}\boldsymbol{y} \leqslant b\}$表示输出向量 $\boldsymbol{y}$ 的集合，该集合会产生观测到的活动集和在 λ 处 lasso 所选择的预测子符号。这些不等式是从次梯度条件推导得出的（见习题 6.10）。这会产生一组有用的变量集 $\mathcal{A}$，例如，现在可以对 $\boldsymbol{y}$ 在 $\boldsymbol{X}_{\mathcal{A}}$ 上的总体回归系数做条件推断。这意味着要对 $\boldsymbol{\eta}$ 等于 $\boldsymbol{X}_{\mathcal{A}}(\boldsymbol{X}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}})^{-1}$ 的每一列做单独的条件分析。因此在 λ 处 lasso 的解中，可以得到活动集参数的精确 p 值和置信区间。这些值有正确的 I 型错误率，以及关于成员和活动集[①]符号的覆盖条件。

图 6-12 给定 λ=7，用 lasso 对糖尿病数据得出推断结果，有 7 个变量被选中。注意，现在仅用包含 7 个预测子的降维模型拟合 OLS 回归系数。蓝色区间基于常见的多元回归正态理论，忽略了使用数据是为了从全部 64 个变量中选择出 7 个变

① Lee, Sun, Sun and Taylor (2013) 也讨论了不以符号为条件的推断，通过考虑有相同活动集的所有区域的联合。

量的事实。红色的后选择区间由反关系式（6.10）构成，并且考虑到了选择。可以看到，这两组区间对于大的系数是相似的，但对于小的系数而言，选择调整后会更长一些。

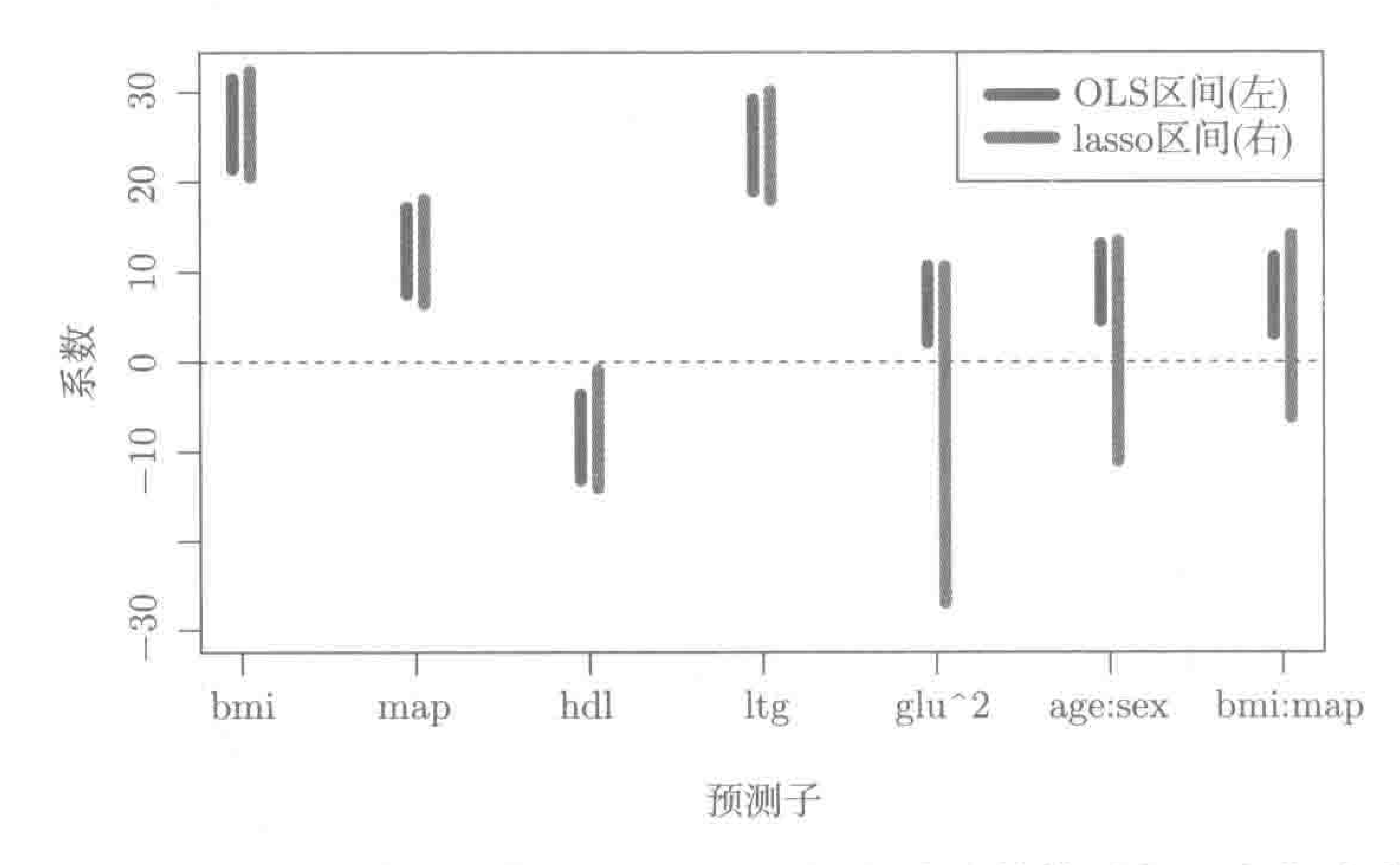


图 6-12　用 lasso 模型拟合糖尿病数据。当 $\lambda=7$ 时，解生成的模型有 7 个非零系数。图为用选择出的变量做最小二乘拟合时的 95% 置信区间。OSL 区间忽略了选择，而 lasso 区间在高斯分布假设下很精确，并以选择事件为条件。其中 λ 由交叉验证（1SE 法则）选出，式（6.10）中的 σ^2 则来自所有 64 个变量的全回归的残差估计

如何选出 $\lambda=7$? 这里使了点小技巧，采用了十重交叉验证（运用了一个标准差法则）。实际应用需要以选择事件为条件，但这将极大增加选择集的复杂程度。模拟显示，这并没有显著地扩大置信区间。下面会讨论在 LAR 序列$\{\lambda_k\}$上的条件推断，这将在 lasso 解路径上限制 λs 集合到节点上。式（6.10）中的 σ 也需要估计。因为 $N>p$，所以可以用 64 个预测子上的全回归的根均方误差来估计。

2. LAR 间距检验

这里将推断过程（6.10）用到一系列 LAR 算法步骤中。例 6.1 的第一步检验了全局空假设。这里设 $\boldsymbol{\eta}_1^{\mathrm{T}}\boldsymbol{y}=\lambda_1=\max_j|\langle\boldsymbol{x}_j,\boldsymbol{y}\rangle|$，检验意味着测试最大协方差是否会偶然超过了预期。$\mathcal{V}^-=\lambda_2$，$\mathcal{V}^+=+\infty$，因此检验可以写成非常简单的形式

$$R_1=1-F_{\mathbf{0},\sigma^2}^{\mathcal{V}^-,\mathcal{V}^+}(\lambda_1|\{\boldsymbol{A}\boldsymbol{y}\leqslant b\})=\frac{1-\Phi(\lambda_1/\sigma)}{1-\Phi(\lambda_2/\sigma)}\sim U(0,1) \tag{6.11}$$

显然，上面的均匀分布对有限的 N 和 p 及任意 $\boldsymbol{X}$ 均完全成立。这就是对全局空假设的**间距**检验（Taylor et al. 2014），协方差检验的非渐近版本，与之渐近相等（见习题 6.5）。间距检验是 $\lambda_1-\lambda_2$ 的单调函数，间距越大，p 值越小。

同样，对于给定 LAR 步骤中增加的变量的部分回归系数是否为零，有更一般的间距检验形式。这些检验基于 λ_k 的连续值，得到的公式比式（6.11）更加复杂。

具体而言，如果变量 $\boldsymbol{x}_{jk}$ 是在 LAR 算法的第 k 步选择的，可以证明对应的节点 λ_k 能由 $\lambda_k = \boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{y}$ 得到，其中

$$\boldsymbol{\eta}_k = \frac{\boldsymbol{P}_{\mathcal{A}_{k-1}}^{\perp}\boldsymbol{x}_{jk}}{s_k - \boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}_{k-1}}(\boldsymbol{X}_{\mathcal{A}_{k-1}}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}_{k-1}})^{-1}s_{\mathcal{A}_{k-1}}} \tag{6.12}$$

（见习题 6.8）。这里 $\mathcal{A}_{k-1}$ 表示 $k-1$ 步后的活动集，且

$$\boldsymbol{P}_{\mathcal{A}_{k-1}}^{\perp} = \boldsymbol{I}_N - \boldsymbol{X}_{\mathcal{A}_{k-1}}\left(\boldsymbol{X}_{\mathcal{A}_{k-1}}^{\mathrm{T}}\boldsymbol{X}_{\mathcal{A}_{k-1}}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_{k-1}}$$

是残差投影算子，用于“调整”$\boldsymbol{X}_{\mathcal{A}_{k-1}}$ 的 $\boldsymbol{x}_{jk}$。最后，s_k 和 $s_{\mathcal{A}_{k-1}}$ 是变量 k 系数的符号，这些是按照 $\mathcal{A}_{k-1}$ 来排序的（后面的那项是一个$(k-1)$ 维向量）。使用 $\boldsymbol{\eta}$ 值，间距检验从上面所列的一般推断过程开始，终止于式（6.10）。节点 λ_k 处的矩阵 $\boldsymbol{A}$ 相比固定 λ 的情况行数更多，因为这里以整个序列 $\{\lambda_\ell\}_1^k$ 为条件。然而计算是可控的，这里可以针对给定 λ 时所选择出的变量精确计算 p 值和置信区间。

Taylor et al.（2014）给出了间距检验的简单版本，即计算精确值的近似，在实验中十分接近，渐近相等（协方差检验也是如此）。这些版本中最值得注意的是

$$R_k = \frac{\Phi\left(\dfrac{\lambda_{k-1}}{\sigma\|\boldsymbol{\eta}_k\|_2}\right) - \Phi\left(\dfrac{\lambda_k}{\sigma\|\boldsymbol{\eta}_k\|_2}\right)}{\Phi\left(\dfrac{\lambda_{k-1}}{\sigma\|\boldsymbol{\eta}_k\|_2}\right) - \Phi\left(\dfrac{\lambda_{k+1}}{\sigma\|\boldsymbol{\eta}_k\|_2}\right)} \tag{6.13}$$

这是使用 $\mathcal{V}^- = \lambda_{k-1}$，$\mathcal{V}^+ = \lambda_{k+1}$ 对式（6.11）的精确泛化。很容易看出，重要的一项 [式（6.13）的右上部] 为

$$\tilde{\theta}_k = \frac{\lambda_k}{\sigma\|\boldsymbol{\eta}\|_2} = \frac{\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{y}}{\sigma\|\boldsymbol{\eta}\|_2} \tag{6.14}$$

这是 $\boldsymbol{X}_{\mathcal{A}_{k-1}}$ 中 $\boldsymbol{x}_{jk}$ 的（绝对）标准偏回归系数（见习题 6.9）。这说明 λ_k 的检验等价于偏回归系数的检验。

表 6-2 最右列是在糖尿病数据上采用这种更一般的间距检验所得到的结果。这与协方差检验的结果相似。

尽管间距检验和固定 λ 值方法在构建方面很相似，而且精确性也一样，但在某个重要方面却不一样。具体而言，间距检验适用于连续 LAR 过程中的每一步，并且会使用具体的 λ 值（节点）。而固定 λ 值的推断可以用在任意 λ 值上，但选中之后就视为固定值。因此该方法忽略了从数据中选择 λ 所带来的任意附加可变性。

6.3.3 检验何种假设

在自适应检验中，这个问题很棘手。协方差检验使用了一组条件假设：LAR 的每一步会检验所有不在模型中的其他预测子的系数是否为零。有时候称之为**完全空假设**。

事实证明，间距检验有不同的侧重点。在第一步，同协方差检验一样，算法检验全局空假设。但在接下来的步骤，算法会检验该步所选择的预测子的偏相关系数是否为零，以调整当前模型中的其他变量。这有时也称为**增量空**假设。同协方差检验不同，算法无意估计当前模型的全局正确性。固定 λ 检验与之类似，算法以预测子的当前活动集为条件，并检验在投影模型中任意给定预测子的系数是否为零。6.4 节会讨论另一种方法，用于计算全模型中总体回归系数的置信区间。

6.3.4　回到向前逐步回归

前面谈到过，向前逐步回归的朴素推断忽略了选择的影响，如图 6-8a 和表 6-2 左侧所示。在介绍完这些之后，读者可能注意到，6.3.2 节的普通推断算法事实上可以用在向前逐步回归上，这也为该过程提供了合适的选择推断。在这种情况下，约束矩阵 $\boldsymbol{A}$ 稍微有点复杂，在第 k 步大约有 $2pk$ 行。但其计算复杂性是可接受的，详见 Taylor et al.（2014）以及 Loftus and Taylor（2014）。

6.4　通过去偏 lasso 推断

这里介绍的方法和 6.3 节讨论的方法在目标上大不一样。这个方法无意推断 LAR 或者 lasso 生成的模型的偏回归系数，而是在假设的线性模型下，直接估计整个总体回归系数集的置信区间。为此，该方法将 lasso 估计 [①] 作为开端，用去偏操作产生用于构建置信区间的估计。

假设线性模型 $\boldsymbol{y}=\boldsymbol{X}\beta+\boldsymbol{\varepsilon}$ 是正确的，现在需要 $\{\beta_j\}_1^p$ 的置信区间。如果 $N>p$，可以简单地通过最小二乘法拟合全模型，并用最小二乘理论中的标准区间

$$\hat{\beta}_j \pm z^{(\alpha)} v_j \hat{\sigma} \tag{6.15}$$

其中 $\widehat{\beta}$ 是 OLS 估计，$v_j^2=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}_{jj}$，$\widehat{\sigma}^2=\sum_i(y_i-\widehat{y}_i)^2\Big/(N-P)$，$z^{\alpha}$ 是标准正态分布的 α 分位数。但是这个方法在 $N<p$ 的时候不能使用。

之前有人（Zhang and Zhang 2014, Bühlmann 2013、van de Geer, Bühlmann, Ritov and Dezeure 2013、Javanmard and Montanari 2014）提了一种方案，使用 lasso 估计的去偏版本，即

$$\hat{\beta}^d=\hat{\beta}_\lambda+\frac{1}{N}\boldsymbol{\Theta}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}_\lambda\right) \tag{6.16}$$

其中 $\widehat{\beta}_\lambda$ 是 λ 处的 lasso 估计，$\boldsymbol{\Theta}$ 是 $\boldsymbol{\Sigma}=\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的近似逆。[②] 因此有

---

① 基于一致性考虑，用 λ 值拟合。

② 如果 $N\geqslant p$，则 $\boldsymbol{\Theta}^{-1}=\dfrac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$，式 (6.16) 即是 β 的精确无偏估计。但是当 $N<p$，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/N$ 不可逆，则需要寻找一个近似逆。

$$\hat{\beta}^d = \beta + \frac{1}{N}\boldsymbol{\Theta}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} + \underbrace{\left(\boldsymbol{I}_P - \frac{1}{N}\boldsymbol{\Theta}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\hat{\beta}_\lambda - \beta\right)}_{\hat{\Delta}} \tag{6.17}$$

其中 $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_p\right)$。这些作者提供了 $\boldsymbol{\Theta}$ 的（不同）估计，使 $\|\widehat{\Delta}\|_\infty \to 0$。对于式 (6.17)，可以用近似 $\widehat{\beta}^d \sim N(\beta, \frac{\sigma^2}{N}\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}^{\mathrm{T}})$ 来计算 β_j 的置信区间。去偏算子 (6.16) 可以看作优化残差平方和的近似牛顿步骤，从 lasso 估计 β 开始（见习题 6.6）。对于估计 $\boldsymbol{\Theta}$，另有一些建议如下。

- van de Geer et al. (2013) 采用基于近邻的方法来估计 $\boldsymbol{\Theta}$，以便在元素上产生稀疏性（详见第 9 章）。
- Javanmard and Montanari (2014) 用了另外一种方法：对每一个 j 定义 m_j，以求解凸问题

$$\underset{m \in \mathbb{R}^p}{\text{minimize}} \quad m^T \hat{\Sigma} m，其约束为 \tag{6.18}$$

$$\left\|\hat{\Sigma} m - e_j\right\|_\infty \leqslant \gamma \tag{6.19}$$

其中 e_j 是第 j 个单位向量。然后他们设

$$\widehat{\boldsymbol{\Theta}} := (m_1, m_2, ..., m_p) \tag{6.20}$$

即要求 $\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}} \approx \boldsymbol{I}$，且 $\widehat{\beta}_j^d$ 的方差变小。

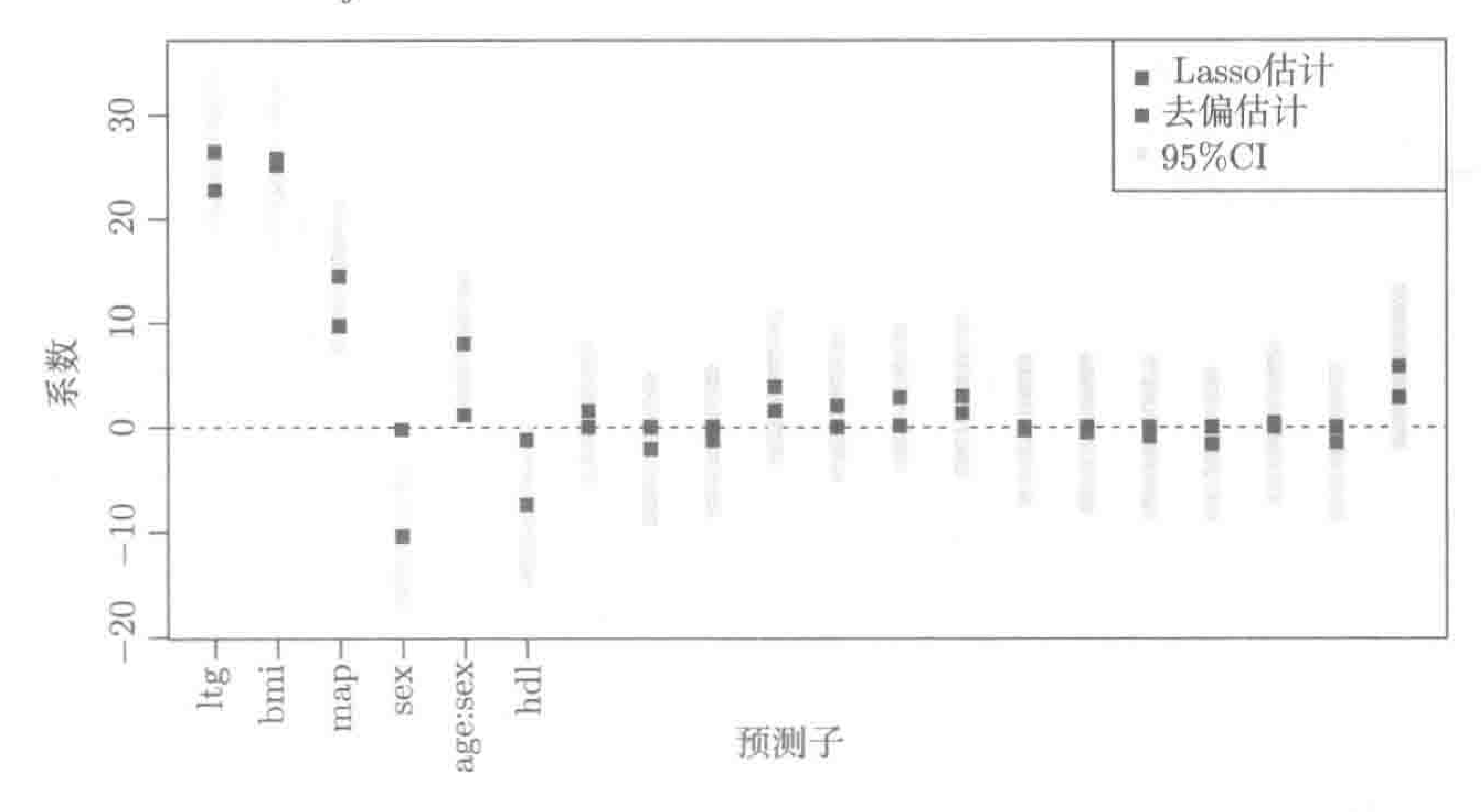


图 6-13 糖尿病数据：lasso 估计、去偏 lasso 估计以及去偏方法的置信区间。这些区间并没有针对多重比较来调整。前 6 个预测子的区间并没有包含 0。进行 Bonferroni 调整之后，这个数目降低到 3

图 6-13 是采用 Javanmar and Montanari (2014) 的去偏方法在糖尿病数据上得到的结果。其中 6 个预测子的 95% 的置信区间不包含 0。但是这些区间并没有针

对多重比较校正过。如果用 0.05/64 的 Bonferroni 调整，则显著预测子的数目降低到 3 个。最显著的 3 个预测子与表 6-2 中协方差检验和间距检验得到的相符，第四个预测子（性别）直到表 6-2 中向前逐步算法的第 7 步才被选入，在其他两种方法的前十步没有出现。

6.5　后选择推断的其他建议

PoSI 方法（Berk, Brown, Buja, Zhang and Zhao 2013, Post Selection Inference, 后选择推断）用来拟合选择后的子模型，然后通过选择方法得到所有可能的模型，并用这些模型来调整标准（非自适应）置信区间。这种调整不用于寻找给定模型的搜索方法。这样做既有优点也有缺点。优点是，用方法可针对非作者指定的搜索方法产生结果，或者置疑提出的算法是否考虑到实际的应用情况。缺点在于，算法为了体现其健壮性产生了相当宽（保守）的置信区间。

这里再一次考虑线性模型 $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$，假设模型选择算法 $\mathcal{M}$ 选择了一个子模型 M，估计为 $\widehat{\beta}_M$。PoSI 的提出者认为，推断不应当关乎真正内在的参数向量 $\boldsymbol{\beta}$，而是关乎 $\boldsymbol{X}\beta$ 在 $\boldsymbol{X}_M$ 上投影的参数

$$\beta_M = \left(\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M\right)^{-1}\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}\beta \tag{6.21}$$

这个方法也被 6.3.2 节中所讨论的条件推断所采用。考虑 β_M 的第 j 个元素的置信区间形为

$$\mathrm{CI}_{j\cdot M} = \hat{\beta}_{j\cdot M} \pm K\hat{\sigma}v_{j\cdot M} \tag{6.22}$$

其中 $v_{j\cdot M}^2 = (\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M)_{jj}^{-1}$。然后 PoSI 算法给出一个常数 K，使

$$\Pr\left(\beta_{j\cdot M} \in \mathrm{CI}_{j\cdot M}\right) \geqslant 1 - 2\alpha \tag{6.23}$$

覆盖所有可能的模型选择算法 $\mathcal{M}$。K 值是数据矩阵 $\boldsymbol{X}$ 的函数，也是 β_M 中非零元素（而非输出向量 $\boldsymbol{y}$）的最大数目。这里要证明，在正交情况下 K 以 $\sqrt{2\log(p)}$ 的速度增长，在非正交情况下以 $\sqrt{p}$ 的速度增长。

注意，式 (6.21) 的投影子模型中任意单个参数可以写成 $a^{\mathrm{T}}\beta$，最小二乘估计为 $a^{\mathrm{T}}\widehat{\beta}$，其中 $\widehat{\beta}$ 是全模型的最小二乘估计。Scheffe (1953) 为所有这类线性组合提供一种算法，得到共同推断：

$$\Pr\left[\sup_a \frac{\left[a^{\mathrm{T}}(\hat{\beta}-\beta)\right]^2}{a^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}a\cdot\hat{\sigma}^2} \leqslant K_{\mathrm{Sch}}^2\right] = 1 - 2\alpha \tag{6.24}$$

假设有高斯误差的全模型是正确的，可以证明 $K_{\mathrm{Sch}} = \sqrt{pF_{p,N-p,1-2\alpha}}$，它能得到前面提到的上界 $\sqrt{p}$。PoSI 的提出者证明：采用数值方法和直接检索，对真实模型矩阵 $\boldsymbol{X}$，能找到较小的 K 值，特别是在某个最大尺寸中用户限定所有模型时。

对糖尿病数据，Andreas Buja 计算除了维度为 5 的子模型的 K 值（计算时间将近 2 小时）。得到的 K 值为 4.21（90%）、4.42（95%）和 4.85（99%）。在 95% 水平下，得出 4 个显著预测子 bmi、map、hdl 和 ltg。如果后者区间针对多重比较进行了调整，则会比按图 6-12 的 lasso 方法要多得到一个预测子。

与图 6-12 中相比，PoSI 区间更有优势，因为不需要知道 σ 和固定 λ 值。另一方面，PoSI 的置信区间可以十分宽。在糖尿病数据中，有 4 个显著预测子，lasso 区间本质上没有受选择方法所影响，看起来仍像是标准最小二乘区间。即使进行从 0.05 到 0.01 的 Bonferroni 调整，区间近似长度为 $\pm 2.23 \cdot \sigma v_{j \cdot M}$，而 PoSI 的则为 $\pm 4.42 \cdot \sigma v_{j \cdot M}$。但是 PoSI 的提出者认为：他们的方法对处理数据时所做的所有（未报导的）事情提供了较强的保护，例如获取拥有大量显著预测子的模型。

PoSI 最主要的缺点就是计算能力有限。正如提出者所说，并行计算能够求解数据维度 64 下模型维度 7 或者 8 的模型，这就是极限了。

参考文献注释

本章讨论的贝叶斯 lasso 由 Park and Casella（2008）提出。自助法源于 Efron（1979），Efron and Tibshirani（1993）是一本综合参考书。贝叶斯方法和自助法之间的联系在很多文献中探讨过（Rubin 1981, Efron 1982, Efron 2011）。

协方差检验是由 Lockhart, Taylor, Tibshirani_2 and Tibshirani（2014）引入的。这里的讨论就基于这篇论文，该论文是模型选择方面的实用资源。这个工作可以扩展到广义模型和 Taylor, Loftus and Tibshirani_2（2013）中的精确检验。间距检验由 Taylor et al.（2014）提出，Lee, Sun, Sun and Taylor（2013）为 lasso 推导了固定 λ 推断算法。Taylor et al.（2014）和 Loftus and Taylor（2014）提出了向前逐步回归检验，后者包括了组 lasso 惩罚的分类变量。Grazier G'Sell, Wager, Chouldechova and Tibshirani（2015）为序列检验提出了 FDR 控制算法，并且将其运用到模型选择 p 值上。Grazier G'Sell Taylor and Tibshirani（2013）对图 lasso 算法给出了协方差检验，Choi, Taylor 和 Tibshirani（2014）对主成分进行了同样的处理。Fithian, Sun and Taylor（2014）着重于指数族分布，为模型选择后的条件推断提供了一种广义理论框架。

去偏方法（6.4 节）时有提及。比如，Zhang and Zhang（2014）针对高维回归系数对比，用松弛投影中的残差（例如，稀疏线性回归中的偏差）替换常见的评分向量，从而推导出了置信区间。Bühlmann（2013）计算了高维回归模型中系数的 p 值，从岭估计开始，然后得到使用 lasso 的偏差校正项。van de Geer et al.（2013），Javanmard and Montanari（2014）和 Javanmard and Montanari（2013）继续了这项开创性的工作，提出了很多方法。这些方法都基于预测子的逆方差矩阵的估计来进

行去偏 lasso 估计。(后者的工作集中在特例上，如预测子矩阵 $\boldsymbol{X}$ 的行服从独立同分布的高斯分布；其他人考虑的则是常见矩阵$\boldsymbol{X}$ 。) 这些去偏 lasso 估计是渐近正态的，可以计算单个系数的边际 p 值，也可同时计算一组系数的共同 p 值。PoSI 算法由 Berk et al.（2013）提出。

习　　题

习题 6.1　(a) 求证：对于正交情况 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}_{p\times p}$, 对所有的步骤 k 协方差检验式（6.5）可简化为

$$T_k = \frac{1}{\sigma^2} \cdot \lambda_k \left(\lambda_k - \lambda_{k+1}\right) \tag{6.25}$$

(b) 求证：对于一般的 $\boldsymbol{X}$，第一步（$k = 1$）的协方差检验式（6.5）可以简化为式（6.25）。

习题 6.2　求证：式（6.4）中的 R_k 可以写成方差统计量

$$R_k = \frac{1}{\sigma^2} \cdot \left(\left\langle \boldsymbol{y}, \boldsymbol{X}\hat{\beta}_k \right\rangle - \left\langle \boldsymbol{y}, \boldsymbol{X}\hat{\beta}_{k-1} \right\rangle\right) \tag{6.26}$$

其中 $\widehat{\beta}_k$ 是向前逐步回归在第 k 步后的系数向量。(这些变量的系数不包括设为 0 的部分。)

习题 6.3　FDR 顺序控制。假设用 p 值 $p_1, p_2, \ldots, p_m$ 进行一组假设检验 $H_0^1, H_0^2, \ldots, H_0^m$。将 p 值排序为 $p_{(1)} < p_{(2)} < \ldots < p_{(m)}$。假设有一算法拒绝了所有假设中的 R 个，且有 V 个为假阳性，则定义算法的假阳性率为 $\mathrm{E}(V/R)$。给定一个目标 FDR 为 α，Benjamini–Hochberg（BH）算法（Benjamini and Hochberg 1995）拒绝了有最小 p 值的 R 个假设，其中 R 是使得 $p_{(j)} \leqslant \alpha \cdot j/m$ 的最大值。如果 p 值是独立的，则这个算法的 FDR 的最大值为 α。

(a) 计算糖尿病数据中各个预测子的单变量回归系数 $\widehat{\beta}_j$ 和标准差 $\widehat{\mathrm{se}}_j$。可以得到近似的正态分数值 $z_j = \widehat{\beta}_j \big/ \widehat{\mathrm{se}}_j$，以及相关的（二）截尾 p 值。运用 BH 算法寻找 FDR 为 5%的一组显著预测子。

(b) 假设必须按顺序考虑问题，即拒绝一组连续的 K 个假设 $H_0^1, H_0^2, \ldots, H_0^K$（或者一个也不拒绝）。协方差检验或间距检验即是这样的例子。BH 算法在这种情况下无法使用，因为它没有考虑排序。例如在表 6-2 中，BH 算法可能拒绝`ltg`的空假设，但是并没有拒绝`bmi`。这是无用的，因为要找的是包含前 k 个预测子（对一些 $k \geqslant 0$）的模型。这里可使用 BH 算法的推广形式。设协方差检验或者间距检验的 p 值为 $p_1, p_2, \ldots, p_m$，$r_k = -\sum_{j=1}^{k} \log\left(1-p_j\right)/k$。所谓的前停（ForwardStop）准则拒绝了 $p_1, p_2, \ldots,$

$p_{\widehat{k}}$，其中 $\widehat{k}$ 是使得 $r_k \leqslant \alpha$ 的最大 k 值（Grazier G'Sell et al. 2015）。将前停准则用到协方差检验或者间距检验 p 值上，其目标 FDR 为 5%。

习题 6.4　这里推导一个有关多维正态分布的定理，在 (c) 中，用该定理来推导在全局空假设下的 LAR 间距检验。假设随机向量 $Z=(Z_1,\ldots,Z_p)$ 服从多维正态分布 $N(0,\boldsymbol{\Sigma})$，对所有的 j 有 $\boldsymbol{\Sigma}_{jj}=1$。

(a) 设

$$(j_1,s_1)=\underset{j\in\{1,2,\ldots,p\},s\in\{-1,1\}}{\arg\max}(sZ_j)$$

假设这些索引是单独得到的。定义随机变量

$$M_j=\max_{1\leqslant i\leqslant p,i\neq j,s\in\{-1,1\}}\left\{\frac{sZ_i-s\Sigma_{ij}Z_j}{1-ss_j\Sigma_{ij}}\right\} \tag{6.27}$$

其中 $s_j=\arg\max_{s\in\{-1,1\}}(sZ_j)$。求证：对于所有的 $j=1,2,\ldots p$，M_j 与 Z_j 独立。

(b) 设 $\Phi(x)$ 为标准高斯分布的 CDF，且

$$U(z,m)=\frac{1-\Phi(z)}{1-\Phi(m)} \tag{6.28}$$

验证当且仅当 $Z_j\geqslant M_j$ 有 $j_1=j$，并证明 $U(Z_{j1},M_{j1})$ 在 (0,1) 上均匀分布。

(c) 在 LAR 算法中将预测子归一化，设 $\sum=\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$，$Z_j=\frac{1}{N}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y}$。求证：$\lambda_1=\max_{j,s}(sZ_j)$，$\lambda_2=M_{j1}$（难），并推导间距检验（6.11）。

习题 6.5　求证：随着 $N,p\to\infty$，协方差检验式（6.5）和间距检验式（6.11）渐近相等。提示：设 λ_2 以一定速率趋向于无穷，所以 $\lambda_1/\lambda_2\to\infty$，可以采用 Mill 比率。

习题 6.6　考虑残差平方和函数 $J(\beta)=\|\boldsymbol{y}-\boldsymbol{X}\beta\|^2$，构建最小化 $J(\beta)$ 的牛顿步骤形为

$$\beta^{\text{new}}\leftarrow\beta+\left(\frac{\partial J}{\partial\beta}\right)^{-1}\frac{\partial J}{\partial\beta} \tag{6.29}$$

其中 β 是在 λ 处的 lasso 估计。证明此式有式（6.16）的形式，其中用 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 替换为式（6.20）中的估计 $\widehat{\boldsymbol{\Theta}}$。

习题 6.7　LAR 算法和 lasso 的一般推断。设 $\boldsymbol{y}\sim N(\boldsymbol{\mu},\sigma^2\boldsymbol{I})$，考虑选择事件 $\{\boldsymbol{A}\boldsymbol{y}\leqslant b\}$ 下 $\boldsymbol{y}$ 的分布。

(a) 求证：

$$\{\boldsymbol{A}\boldsymbol{y}\leqslant b\}=\left\{\mathcal{V}^-(\boldsymbol{y})\leqslant\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}\leqslant\mathcal{V}^+(\boldsymbol{y}),\nu^0(\boldsymbol{y})\geqslant 0\right\} \tag{6.30}$$

其中的变量定义为

$$\begin{aligned}\alpha &= \frac{\boldsymbol{A\eta}}{\|\boldsymbol{\eta}\|^2}\\ \mathcal{V}^-(\boldsymbol{y}) &= \max_{j:\alpha_j<0}\frac{b_j-(\boldsymbol{Ay})_j+\alpha_j\boldsymbol{\eta}^T\boldsymbol{y}}{\alpha_j}\\ \mathcal{V}^+(\boldsymbol{y}) &= \min_{j:\alpha_j>0}\frac{b_j-(\boldsymbol{A}y)_j+\alpha_j\boldsymbol{\eta}^T\boldsymbol{y}}{\alpha_j}\\ \mathcal{V}^0(\boldsymbol{y}) &= \min_{j:\alpha_j=0}\Big(b_j-(\boldsymbol{A}y)_j\Big)\end{aligned} \tag{6.31}$$

提示：从不等式 $\boldsymbol{Ay}\leqslant b$ 两边同时减去 $\mathrm{E}\left(\boldsymbol{Ay}|\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y}\right)$；分别简化和检验 $\alpha_j<0,=0$ 和 >0 的情况。

(b) 设

$$F^{c,d}_{\mu,\sigma^2}(x)=\frac{\Phi\left((x-\mu)/\sigma\right)-\Phi\left((c-\mu)/\sigma\right)}{\Phi\left((d-\mu)/\sigma\right)-\Phi\left((c-\mu)/\sigma\right)} \tag{6.32}$$

这是一个截尾正态分布，区间为 $[c,d]$。求证：

$$F^{\mathcal{V}^-,\mathcal{V}^+}_{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu},\sigma^2\|\boldsymbol{\eta}\|^2}(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{y})\,|\{\boldsymbol{Ay}\leqslant b\}\sim U(0,1) \tag{6.33}$$

在 LAR 算法的给定步骤或者固定 λ 值计算出的 lasso 解中，这个结果可以用来推断参数 $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu}$。

(c) 基于结果（6.33），给出间距检验结果（6.11）的另一种证明。

习题 6.8　LAR 算法中的第 k 个节点是 λ_k 值，在这里第 k 个变量进入模型。在 λ_k 处该变量的系数为 0（即将从 0 开始增长）。运用 KKT 优化条件，验证表达式（6.12）。

习题 6.9　用式（6.12）中定义的 $\boldsymbol{\eta}_k$，证明式（6.14）中的 $\widetilde{\theta}_k$ 是 $\boldsymbol{y}$ 在 $\boldsymbol{x}_{jk}$ 上的绝对归一化偏回归系数对 $\boldsymbol{X}_{\mathcal{A}_{k-1}}$ 的调整。

习题 6.10　考虑 lasso 问题

$$\underset{\beta}{\text{minimize}}\,\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2+\lambda\|\beta\|_1$$

令 $E\subset\{1,\ldots,p\}$ 表示一个候选活动集，$s_E\in\{-1,1\}^{|E|}$ 为活动变量的符号。有相同 E 和 S_E 的任意解 $\widehat{\beta}_E$ 对应的 KKT 条件为

$$-\boldsymbol{X}_E^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}_E\hat{\boldsymbol{\beta}}_E)+\lambda s_E=0 \tag{6.34}$$

$$-\boldsymbol{X}_{-E}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}_E\hat{\boldsymbol{\beta}}_E)+\lambda s_{-E}=0 \tag{6.35}$$

其中 $\mathrm{sgn}(\widehat{\beta}_E)=s_E$，$\|s_{-E}\|_\infty<1$。消除这些等式中的 $\widehat{\beta}_E$，求证 $\boldsymbol{y}$ 的值和解 $(E,\ s_E)$ 可以通过一中组线性不等式 $\boldsymbol{Ay}\leqslant b$ 来定义。

习题 6.11　考虑例 6.1 的设定，假设 $\boldsymbol{x}_{j_1}^{\mathrm{T}}\boldsymbol{y}$ 为正。采用简单不等式，推导 $\mathcal{V}^-(\boldsymbol{y})$ 的表达式。求证该式等于 λ_2（第 2 个 LAR 节点）。

第 7 章　矩阵的分解、近似及填充

7.1 引　　言

本章关注的问题是：给定数据 $m \times n$ 矩阵 $\boldsymbol{Z} = \{z_{ij}\}$，按某种意义找到 $\boldsymbol{Z}$ 的近似 $\widehat{\boldsymbol{Z}}$。其目的是通过具有简单结构的近似 $\widehat{\boldsymbol{Z}}$ 来理解 $\boldsymbol{Z}$。另一个目的是填充 $\boldsymbol{Z}$ 中缺失的元素，这个问题称为**矩阵填充**（matrix completion）。

通常的方法是基于形如

$$\hat{\boldsymbol{Z}} = \arg \min_{\boldsymbol{M} \in \mathbb{R}^{m \times n}} \|\boldsymbol{Z} - \boldsymbol{M}\|_F^2, \quad \text{其约束为 } \Phi(\boldsymbol{M}) \leqslant c \tag{7.1}$$

的优化问题来进行估计，其中 $\|\cdot\|_F^2$ 为矩阵的 Frobenius 范数（矩阵中每个元素的平方和）的平方，$\Phi(\cdot)$ 是约束函数，在某些情形下可通过该函数让 $\widehat{\boldsymbol{Z}}$ 稀疏。在此产生稀疏的方式会引出一系列有用的算法，其中有许多会在本章介绍。可以对整个 $\widehat{\boldsymbol{Z}}$ 进行正则化，或者对 $\widehat{\boldsymbol{Z}}$ 进行分解并对分解后的各个部分进行正则化。这里还需注意一些情况，比如 $\boldsymbol{Z}$ 中如果有些元素缺失，那么对范数 $\|\cdot\|_F^2$ 也要做相应的修改。在某些情形下，近似矩阵 $\widehat{\boldsymbol{Z}}$ 会有多个约束。

表 7-1 对本章要讨论的方法进行了汇总。方法 (a) 对 $\widehat{\boldsymbol{Z}}$ 中的所有元素采用 ℓ_1 范数约束。这种约束会得到原矩阵的软阈值版本，即式（7.1）的优化解形如 $\hat{z}_{ij} = \operatorname{sgn}(z_{ij})\left(|z_{ij} - \gamma|\right)_+$，其中选择标量 $\gamma > 0$ 是为了使 $\sum_{i=1}^{m}\sum_{j=1}^{n}|\hat{z}_{ij}| = c$。这样得到的结果 $\widehat{\boldsymbol{Z}}$ 在稀疏协方差估计中很有用。方法 (b) 限定了矩阵 $\widehat{\boldsymbol{Z}}$ 的秩的范围；也就是说，限定了 $\widehat{\boldsymbol{Z}}$ 的非零奇异值的数量。虽然这种秩约束的矩阵近似问题（7.1）非凸，但很容易通过计算奇异值分解（Singular Value Decomposition，SVD）并用前 k 个特征向量得到该问题的优化解。方法 (c) 将秩约束放宽为**原子范数**约束，即矩阵奇异值之和小于给定值。原子范数是一个凸矩阵函数，(c) 中的问题是凸的，可通过计算 SVD 再阈值化相应的奇异值来求得解。当矩阵存在缺失值时，从方法 (b) 到方法 (c) 的修改很重要。通过这样的修改，可精确求解 (c) 所对应的问题，而通常方法 (b) 求解起来会很困难。方法 (d) 会对 $\widehat{\boldsymbol{Z}}$ 的左奇异向量和右奇异向量进行惩罚，其中正则化函数或惩罚函数 Φ_1 和 Φ_2 可用 ℓ_2 或 ℓ_1 范数。ℓ_1 范数会得到稀疏的奇异向量。对奇异向量的解释很重要的地方，这种稀疏性就很有用。方法 (e) 会直接对 LR 矩阵分解所得到的各个部分进行惩罚，这样做表面上与方法 (d) 相似，但当 Φ_1 和 Φ_2 为 Frobeniu 范数时，更类似于方法 (c)。最后，方法 (f) 会将矩阵分解后相加，并对分解的两部分进行惩罚。

表 7-1　矩阵近似问题的不同公式

约束	对应的方法
(a) $\|\hat{\boldsymbol{Z}}\|_{\ell_1} \leqslant c$	稀疏矩阵近似
(b) $\text{rank}(\hat{\boldsymbol{Z}}) \leqslant c$	奇异值分解
(c) $\|\hat{\boldsymbol{Z}}\|_* \leqslant c$	凸矩阵近似
(d) $\hat{\boldsymbol{Z}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$， $\Phi_1(\boldsymbol{u}_j) \leqslant c_1$，$\Phi_2(\boldsymbol{v}_k) \leqslant c_2$	惩罚 SVD
(e) $\hat{\boldsymbol{Z}} = \boldsymbol{L}\boldsymbol{R}^{\mathrm{T}}$， $\Phi_1(\boldsymbol{L}) \leqslant c_1$，$\Phi_2(\boldsymbol{R}) \leqslant c_2$	最大间隔的矩阵分解
(f) $\hat{\boldsymbol{Z}} = \boldsymbol{L} + \boldsymbol{S}$， $\Phi_1(\boldsymbol{L}) \leqslant c_1$，$\Phi_2(\boldsymbol{R}) \leqslant c_2$	矩阵分解的加法形式

矩阵分解也为构造多元统计分析方法（比如主成分分析、典型相关分析、线性判别法等）的稀疏版本提供了思路。在这种情况下，矩阵 $\boldsymbol{Z}$ 本身不是原始数据，而是来自于原始数据。比如，主成分分析基于样本协方差（或对数据矩阵按列进行中心化），典型相关使用的交叉相乘的矩阵来自两组变量，而聚类采用的是训练集中样本之间的距离。本书会介绍这些多元方法，这些问题的相关方法会在第 8 章讨论。

7.2　奇异值分解

给定 $m \times n$ 矩阵 $\boldsymbol{Z}$（其中 $m \geqslant n$），其**奇异值分解**为

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \tag{7.2}$$

这是数值线性代数中的一种标准分解。有很多算法可以高效地计算这种分解（Golub and Loan 1996）。$\boldsymbol{U}$ 是一个 $m \times n$ 正交矩阵（$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_n$），该矩阵的列 $\boldsymbol{u}_j$ 称为**左奇异向量**；$\boldsymbol{V}$ 是一个 $n \times n$ 正交矩阵（$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}_n$），该矩阵的列 $\boldsymbol{v}_j$ 称为**右奇异向量**。$n \times n$ 矩阵 $\boldsymbol{D}$ 是一个对角矩阵，其对角元素为 $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_n \geqslant 0$，这些元素称为**奇异值**。如果这些奇异值 $\{d_\ell\}_{\ell=1}^n$ 是唯一的，则矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 唯一。如果中心化 $\boldsymbol{Z}$ 的列（每一列为一个变量），则右奇异向量 $\{v_j\}_{j=1}^n$ 称为 $\boldsymbol{Z}$ 的**主成分**（principal component）。因此，由单位向量 $\boldsymbol{v}_1$ 可得到线性组合 $\boldsymbol{s}_1 = \boldsymbol{Z}\boldsymbol{v}_1$。在所有可选择的单位向量中，$\boldsymbol{s}_1$ 的样本方差最大，并称 $\boldsymbol{s}_1$ 为 $\boldsymbol{Z}$ 的**第一主成分**，$\boldsymbol{v}_1$ 为对应**方向向量**。依此类推，$\boldsymbol{s}_2 = \boldsymbol{Z}\boldsymbol{v}_2$ 是**第二主成分**，它在所有 $\boldsymbol{s}_1$ 不相关（uncorrelate）的线性组合中，样本方差最大。这方面的内容可见习题 7.1，8.2.1 节也会有更详细的介绍。

奇异值分解可用来求解秩为 q 的矩阵的近似问题。设 $r \leqslant \text{rank}(\boldsymbol{Z})$，$\boldsymbol{D}_r$ 为对角矩阵，除其前 r 个对角元素以外的其他对角元素均置为 0。这样得到优化问题

$$\underset{\text{rank}(\boldsymbol{M})=r}{\text{mininize}} \|\boldsymbol{Z} - \boldsymbol{M}\|_F \tag{7.3}$$

有一个闭解 $\widehat{\boldsymbol{Z}}_r = \boldsymbol{U}\boldsymbol{D}_r\boldsymbol{V}^{\mathrm{T}}$，这种分解称为 r 秩 SVD（见习题 7.2）。得到的 $\widehat{\boldsymbol{Z}}_r$ 具有某种稀疏性，因为除前 r 个奇异值以外，其他的都被置为 0。8.2.1 节介绍主成分分析时会全面讨论 SVD。

7.3 缺失数据和矩阵填充

如果 $\boldsymbol{Z}$ 中有些元素缺失该怎么办？填充或插入矩阵缺失元素的值通常称为**矩阵填充**（Laurent 2001）。在未知矩阵 $\boldsymbol{Z}$ 上不指定约束条件，则对该矩阵填充显然是一个病态问题，此时通常选择秩约束。协同过滤（collaborative filtering）对推荐系统非常有用，可以将它看作一个低秩矩阵填充问题。

SVD 对求解矩阵填充问题非常有效。具体而言，若 $\boldsymbol{Z}$ 中观察到的元素的索引为一个子集 $\Omega \subset \{1,\ldots,m\}\times\{1,\ldots,n\}$，给定这些观察值，一个自然的方法就是寻找一个最低秩的近似矩阵矩阵 $\widehat{\boldsymbol{Z}}$，该矩阵会对 $\boldsymbol{Z}$ 中观测到的元素进行插值，即

$$\text{minimize}\ \text{rank}(\boldsymbol{M})，其约束为\ m_{ij} = z_{ij}，\ (i,j)\in\Omega \tag{7.4}$$

不同于没有缺失值的情况，这种最小化秩问题的计算量很大（属于 NP 难问题），即使中等规模大小的矩阵一般也无法求解。

另外，通过对观测到的 z_{ij} 进行插值来估计 $\boldsymbol{M}$ 经常会太严格，从而导致过拟和。更好的求解方法是允许所得到的 $\boldsymbol{M}$ 与观测值之间有一定的误差。由此，可得到优化问题

$$\text{minimize}\ \text{rank}(\boldsymbol{M})\ ，其约束为\ \sum_{(i,j)\in\Omega}(z_{ij}-m_{ij})^2 \leqslant \delta \tag{7.5}$$

这等于

$$\underset{\text{rank}(\boldsymbol{M})\leqslant r}{\text{minimize}} \sum_{(i,j)\in\Omega}(z_{ij}-m_{ij})^2 \tag{7.6}$$

简言之，需要找到 $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_r$，其秩最大为 r，这是对有观测值的 $\boldsymbol{Z}$ 的最好近似。$\hat{\boldsymbol{Z}}_r$ 的其他元素可作为 $\boldsymbol{Z}$ 的缺失值。优化问题（7.5）中，不同的 δ 会得到不同的解；而优化问题（7.6）中；不同的 r 也会得到不同的解。它们在这一点很像。

遗憾的是，这两种优化问题都是非凸的，因此通常得不到最优解。不过，可以采用启发式算法来得到局部最优解。比如，在开始时随便猜一些缺失值来填充 $\boldsymbol{Z}$，然后计算式（7.3）中矩阵 $\boldsymbol{Z}$ 的 r 秩 SVD，这会得到对缺失值的新估计。不断重复这个过程，直到收敛为止。最后在 (i,j) 处插入的缺失值为 $\hat{\boldsymbol{Z}}_r$ 的第 i 行、第 j 列对应的元素，这在 Mazumder, Hastie and Tibshirani（2010）中有详细的介绍。7.3.2 节会讨论与这类问题相关的凸优化形式，这些形式都基于原子范数，它们可以得到全局最优解。

7.3.1　Netfix 电影挑战赛

Netflix 电影评分挑战赛是矩阵填充最经典的例子（Bennett and Lanning 2007）。Netflix 是一家在线影片租赁公司。其挑战赛从 2006 开始举行，目的是提高系统向顾客推荐影片的能力。Netflix 提供的数据集有 n =17 770 部影片（每部影片为 1 列），顾客数为 m = 480 189（每个用户为 1 行）。顾客对影片的评分范围是 1~5，其中 1 是最差，5 是最好。在训练集中，数据矩阵非常稀疏，仅有 100 万（1%）个被评过分的元素。这个比赛的目的就是要得出没被评价过的影片的得分，以便更好地向顾客推荐影片。在 2006 年，Netflix 采用的是 Cinematch 算法，该算法在一个大的测试集上得到的均方根误差（Root Mean Square Error，RSME）为 0.9525。这个比赛从 2006 年开始，在那一年，获得第一名的算法将 RSME 降低了至少 10%。在 2009 年，这个比赛的最终获胜者是一群研究人员，他们称为 Bellkor's Pragmatic Chaos，这是三个一起赢得比赛的小组的名称。获胜算法采用了大量统计技术，但同其他参与比赛的算法一样，SVD 起到了至关重要的作用。图 7-1 给出了比赛结束后的排行榜。



NETFLIX

Netflix Prize

COMPLETED

Home　Rules　Leaderboard　Update

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22
14	Gravity	0.8643	9.26	2009-04-22 18:31:32
15	Ces	0.8651	9.18	2009-06-21 19:24:53
16	Invisible Ideas	0.8653	9.15	2009-07-15 15:53:04
17	Just a guy in a garage	0.8662	9.06	2009-05-24 10:02:54
18	J Dennis Su	0.8666	9.02	2009-03-07 17:16:17
19	Craig Carmichael	0.8666	9.02	2009-07-25 16:00:54
20	acmehill	0.8668	9.00	2009-03-21 16:20:50
Progress Prize 2007 - RMSE = 0.8723 - Winning Team: KorBell				
Cinematch score - RMSE = 0.9525				

There are currently 51051 contestants on 41305 teams from 186 different countries.
We have received 44014 valid submissions from 5169 different teams; 0 submissions in the last 24 hours.

Questions about interpreting the leaderboard? Please read this.

图 7-1　Netflix 比赛结束后的排行榜

低秩模型为影片评分提供了很好的思路：用户 i 对第 j 部电影评分的模型形如

$$z_{ij}=\sum_{\ell=1}^{r}c_{i\ell}g_{j\ell}+w_{ij} \tag{7.7}$$

其矩阵形式为 $\boldsymbol{Z}=\boldsymbol{C}\boldsymbol{G}^{\mathrm{T}}+\boldsymbol{W}$，其中，$\boldsymbol{C}\in\mathbb{R}^{m\times r}$，$\boldsymbol{G}\in\mathbb{R}^{n\times r}$。在这种模型中，影片有 r 种风格，每一种风格都对应一群喜欢它的观众。第 i 名顾客对第 ℓ 个群的权重为 $c_{i\ell}$，第 j 部电影属于第 ℓ 种风格的分数为 $g_{j\ell}$。按 ℓ（群或风格）将这些项加到一起，然后增加一些噪声，即可得到总体评分。表 7-2 给出了 10 位用户和 10 部评分最高的影片。

表 7-2 Netflix 公司的电影评分数据摘录。电影的分值为 1（最差）~5（最好）。符号 • 代表缺失值，即这部电影没有用户评分

	《辣身舞》	《拜见岳父大人》	《壮志凌云》	《第六感》	《猫鼠游戏》	《天才一族》	《空中监狱》	《大鱼老爸》	《黑客帝国》	《好人寥寥》
用户1	•	•	•	•	4	•	•	•	•	•
用户2	•	•	3	•	•	•	3	•	•	3
用户3	•	2	•	4	•	•	•	•	2	•
用户4	3	•	•	•	•	•	•	•	•	•
用户5	5	5	•	•	4	•	•	•	•	•
用户6	•	•	•	•	•	2	4	•	•	•
用户7	•	•	5	•	•	•	•	3	•	•
用户8	•	•	•	•	•	2	•	•	•	3
用户9	3	•	•	•	5	•	•	5	•	•
用户10	•	•	•	•	•	•	•	•	•	•

比赛会给出一个“测试集”（probe set），里面含有约 140 万个评分元素。这些元素不是随机抽取的，而是出现的时间晚于大多数。图 7-2 给出了在不同的 SVD 秩下训练集和测试集上的均方根误差，同时也给出了用原子范数作正则化的预测结果，下一节介绍这方面。这里对用于训练的数据矩阵的行和列都进行了中心化（即它们的样本均值都为 0），所要拟和的模型为

$$z_{ij}=\alpha_i+\beta_j+\sum_{\ell=1}^{r}c_{i\ell}g_{j\ell}+w_{ij} \tag{7.8}$$

使用双向 ANOVA 回归模型（对不平衡数据）可以分别估计行和列的均值。

从图 7-2 可以看出，迭代SVD 方法相当有效，但它不能保证各种秩都能找到最优解。在这个例子中，相比正则化解，它存在过拟合倾向。下一节会将这种方法与凸松弛建立起联系，从而让算法有收敛性质。

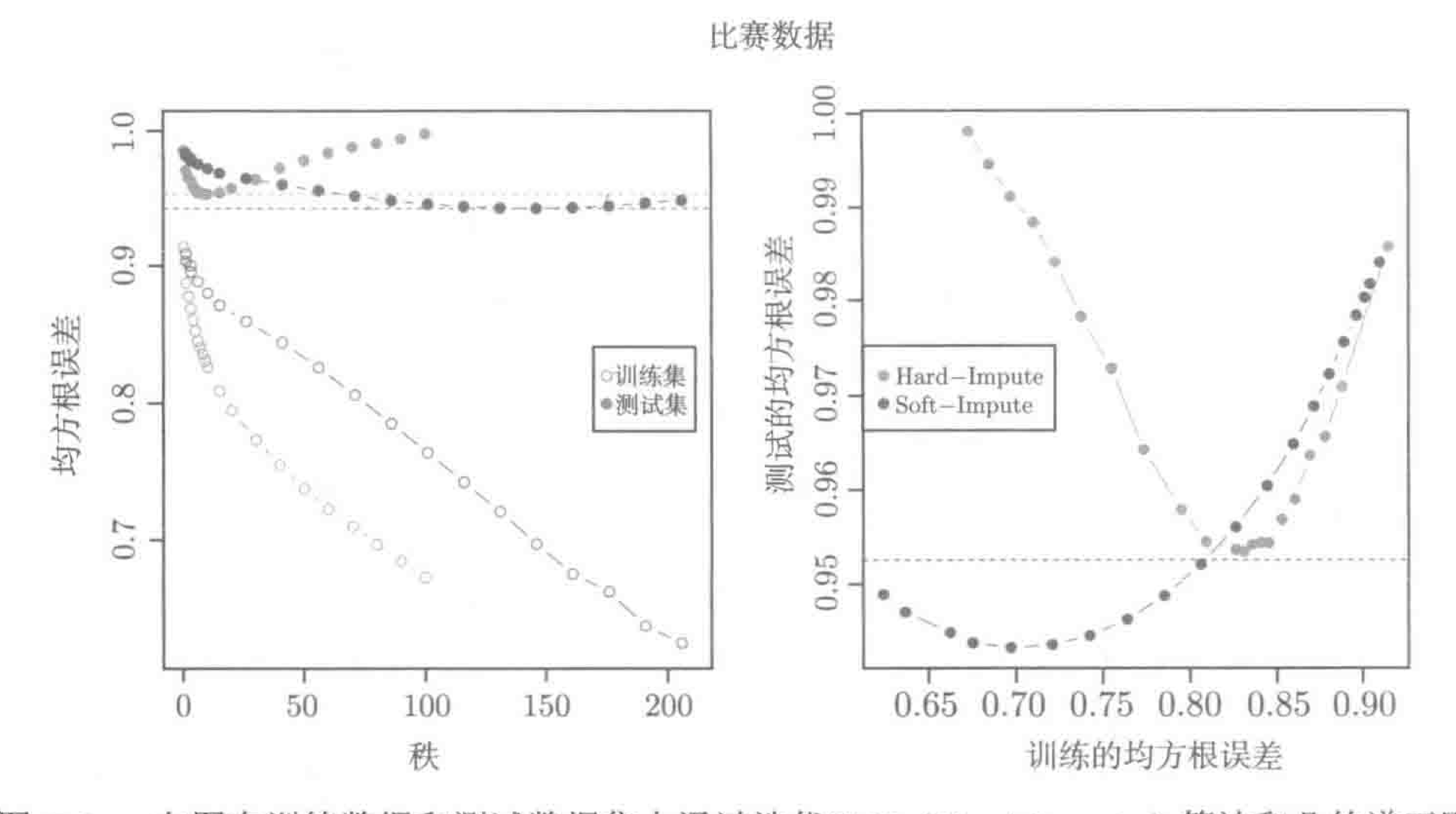


图 7-2　左图在训练数据和测试数据集上通过迭代SVD（Hard-Impute）算法和凸的谱正则化算法（Soft-Impute）得到均方根误差。每个点对应解的秩，这对正则化解来说不是好的度量。右图是两种方法的测试误差，基于训练误差来绘制。训练误差给出了每种方法拟和错误的程度。虚线表示基准方法：Cinematch 评分

7.3.2　基于原子范数的矩阵填充

非凸目标函数（7.4）可松弛成凸形式

$$\text{minimize}\,\|\boldsymbol{M}\|_\star,\quad \text{其约束为 } m_{ij}=z_{ij},\ (i,j)\in\Omega \tag{7.9}$$

其中 $\|\boldsymbol{M}\|_\star$ 为原子范数，即 $\boldsymbol{M}$ 的奇异值之和，有时也称为迹范数 [①]。图 7-3 给出了 2×2 对称矩阵的原子范数的水平集，这也是对凸问题（7.9）的一种形像化描述 [②]。

原子范数是矩阵秩的凸松弛，因此，问题（7.9）是凸优化（Fazel 2002）；更具体地讲，这是半定规划（Semi-Definite Program，SDP，详见习题 7.3）。半定规划是一类特殊的凸优化问题，有具体的求解算法。凸性在理论分析上很有用，因为人们可以给出准确重构矩阵所需要的观测矩阵的性质和样本大小（见 7.3.3 节）。

实际上，由于没有考虑噪声，对观察到的值建立这样的模型并不现实。下面是一种更实际的方法，是式（7.9）的松弛版本：

$$\underset{\boldsymbol{M}}{\text{minimize}}\left\{\frac{1}{2}\sum_{(i,j)\in\Omega}(z_{ij}-m_{ij})^2+\lambda\|\boldsymbol{M}\|_\star\right\} \tag{7.10}$$

① 这个术语可能会造成混淆：对半正定矩阵，迹就是特征值之和。对于一般的矩阵，“迹范数”是指 $\sqrt{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}}$ 的迹，它会将奇异值加在一起。

② 感谢 Emmanuel Candes 和 Benjamin Recht 提供了图 7-3。

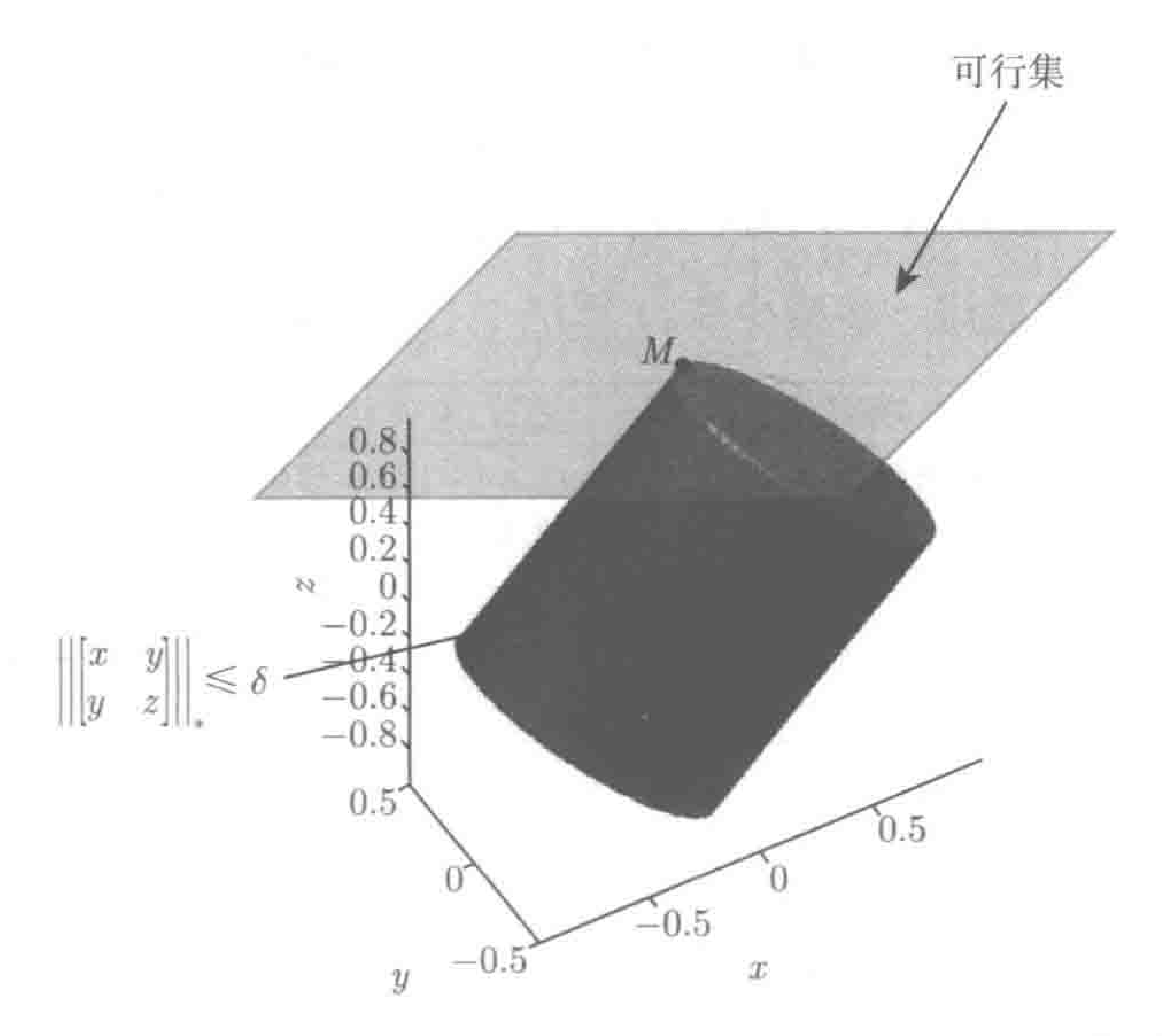


图 7-3 圆柱体表示 2×2 对称矩阵的原子范数单位球水平集的。切平面是矩阵恢复问题中 $z = z_0$ 处的可行集，其中 z 是观测到的值，希望能恢复 x 和 y。点 M 是所要求的解，这会得到最小值 δ。该图类似于 2.2 节中 lasso 估算的图

这称为**谱正则化**（spectral regularization）。前面将式（7.4）松弛成式（7.6），这种修改所得到的解 $\boldsymbol{Z}$ 并不能精确拟和观测值。但在观测值含有噪声的情况下，这种方式可以减少过拟和。参数 λ 要根据数据来进行调整，一般会采用交叉验证来进行调整。和上一节一样，不必让误差 $\sum_{(i,j)\in\Omega}(z_{ij}-m_{ij})^2$ 为 0，仅当 λ 足够小时误差才会为 0。

求解式（7.10）的算法很简单，类似于上一节用来恢复缺失数据的迭代 SVD。首先考虑没有缺失值的情形，即观测到的集合 Ω 有 $m\times n$ 个元素，即 $(i,j)\in\{1,\ldots,m\}\times\{1,\ldots,n\}$。然后求解式（7.10）：计算 $\boldsymbol{Z}$ 的 SVD，然后通过 λ 来软阈值化奇异值，由此重构矩阵。可以采用很简单的方法来设置缺失数据，即先随便猜测缺失值，计算（满秩）SVD，然后通过 λ 来软阈值化奇异值。重构相应 SVD 近似，并获得新缺失值估计，重复该过程直到收敛。

为了更清楚地描述该过程，在此定义几个符号。观测到的元素子集用 Ω 表示，由此定义投影算子 $\mathcal{P}_\Omega:\mathbb{R}^{m\times n}\mapsto\mathbb{R}^{m\times n}$ 为

$$[\mathcal{P}_\Omega(\boldsymbol{Z})]_{ij}=\begin{cases} z_{ij}, & (i,j)\in\Omega \\ 0, & (i,j)\notin\Omega \end{cases} \tag{7.11}$$

即 $\mathcal{P}_\Omega$ 会用 0 替换 $\boldsymbol{Z}$ 的缺失值，只留下观测值。有了这个定义，就可得到等式

$$\sum_{(i,j)\in\Omega}(z_{ij}-m_{ij})^2=\|\mathcal{P}_\Omega(\boldsymbol{Z})-\mathcal{P}_\Omega(\boldsymbol{M})\|_F^2 \tag{7.12}$$

矩阵 $\boldsymbol{W}$ 秩为 r，则相应的奇异值分解 [①] $\boldsymbol{W}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$。这里定义了它的软阈值化版本为

$$S_{\lambda}(\boldsymbol{W}) \equiv \boldsymbol{U}\boldsymbol{D}_{\lambda}\boldsymbol{V}^{\mathrm{T}}\text{ ，其中}\boldsymbol{D}_{\lambda}=\operatorname{diag}\left[(d_1-\lambda)_+,\cdots,(d_r-\lambda)_+\right] \tag{7.13}$$

（注意，软阈值化能进一步减少秩。）算法 7.1 为使用该算子来求解式（7.10）。

算法 7.1 基于 Soft-Impute 的矩阵填充

---

1. 初始化 $\boldsymbol{Z}^{\text{old}}=\mathbf{0}$，创建递减的 $\lambda_1>\ldots>\lambda_k$。
2. 对于每个 $k=1,\ldots,K$，令 $\lambda=\lambda_k$，并进行下面的迭代，直到收敛：
   计算 $\hat{\boldsymbol{Z}}_{\lambda} \leftarrow S_{\lambda}\left(\mathcal{P}_{\Omega}(\boldsymbol{Z})+\mathcal{P}_{\Omega}^{\perp}(\boldsymbol{Z}^{\text{old}})\right)$；
   更新 $\boldsymbol{Z}^{\text{old}} \leftarrow \hat{\boldsymbol{Z}}_{\lambda}$。
3. 输出解序列 $\hat{\boldsymbol{Z}}_{\lambda_1},\ldots,\hat{\boldsymbol{Z}}_{\lambda_K}$。

该算法最早由 Mazumder et al.(2010) 提出，能收敛到全局最优解。习题 7-4 要求验证该算法的不动点会使目标函数（7.10）的次梯度为 0。这个算法与一阶 Nesterov 算法有联系（见习题 7.5）。虽然 $\mathcal{P}_{\Omega}(\boldsymbol{Z})$ 是稀疏的，但每次迭代都需要计算一个（可能很大的）稠密矩阵的 SVD。Netflix 公司提供的数据很大，以至于相应的稠密矩阵通常不能存放在内存中（若矩阵的每个元素占 8 个字节，则需要 68 G 内存）。不过，可以修改为

$$\mathcal{P}_{\Omega}(\boldsymbol{Z})+\mathcal{P}_{\Omega}^{\perp}(\boldsymbol{Z}^{\text{old}})=\underbrace{\mathcal{P}_{\Omega}(\boldsymbol{Z})-\mathcal{P}_{\Omega}^{\perp}(\boldsymbol{Z}^{\text{old}})}_{\text{稀疏}}+\underbrace{\boldsymbol{Z}^{\text{old}}}_{\text{低秩}} \tag{7.14}$$

等式右边第一部分是稀疏的，只有 $|\Omega|$ 个观测值；第二部分是 SVD 的软阈值，所以能够用相应的部分来表示。此外，对于各部分，可以利用其特殊的结构来有效地与向量进行左乘和右乘，从而采用迭代的 Lanczos 方法来有效计算（低秩）SVD。可证明这个迭代算法能收敛到问题

$$\underset{\boldsymbol{M}\in\mathbb{R}^{m\times n}}{\operatorname{minimize}}\left\{\frac{1}{2}\left\|\mathcal{P}_{\Omega}(\boldsymbol{Z})-\mathcal{P}_{\Omega}(\boldsymbol{M})\right\|_F^2+\lambda\|\boldsymbol{M}\|_{\star}\right\} \tag{7.15}$$

的解。上式是目标函数（7.10）的另一种写法。

图 7-2 显示了 Soft-Impute 法在 Netflix 数据集上的结果。从这里可以看出正则化的好处，因为它与迭代 SVD 算法 (Hard-Impute) 相比性能更好。这里花费了更长的时间以减少过拟合，并且采用了正则化，可以让解具有更大的秩。算法 7.1 中采用了热启动，使用 R 包 softImpute（Hastie and Mazumder 2013），花了不到 5 小时就计算出了图 7-2 中的各种解。也可参见 7.3.3 节的图 7-5，它针对噪声矩阵填充，给出了 Soft-Impute 算法在各种不同秩和样本大小下的性能。7.3.3 节会更详细讨论图 7-5。

---

① 如果矩阵的秩 $r<\min(m,n)$，假定该矩阵的 SVD 是以截断形式表示，则丢掉为 0 的奇异值，以及相应的左奇异向量和右奇异向量。

Mazumder et al.（2010）已证明，Soft-Impute 算法至少具有次线性收敛率，即经过 $\mathcal{O}(1/\delta)$ 次迭代所得到的解与全局最优解只差 δ。若没有附加结构（如强凸），这就是一阶梯度法（Nemirovsky and Yudin 1983）所能达到的最快的收敛率。有趣的是，简单的一阶方法在某些情形下会以更快的几何速率收敛，即经过 $\mathcal{O}(\log(1/\delta))$ 次迭代后，就能与最优解相差 δ。例如，Agarwal，Negahban and Wainwright（2012a）分析过一个与 Soft-Impute 密切相关的算法，证明在相同的条件下，基于原子范数估计的统计性能好，该一阶算法能得到几何收敛率。

7.3.3 矩阵填充的理论结果

有多种理论成果基于原子范数正则化的矩阵填充。下面从没有噪声的简单情形开始介绍。假设从 $p \times p$ 矩阵中均匀随机地抽取 N 个元素。对于一个 p 维矩阵，秩为 r，当 N 为多大时，就可用原子范数松弛形式 (7.9) 来精确恢复该矩阵？显然，当 $N \geqslant p^2$ 时总是可以。这里要讨论的是 $N \ll p^2$ 时的情形。

首先很容易发现：如果在某些行（或列）中没有观测值，则不可能准确恢复该矩阵，即使它是一个秩为 1 的矩阵。习题 7.8 将证明任何方法（不仅是原子范数松弛法）都需要 $N > p\log p$ 个观测值才可以准确恢复该矩阵，即使是秩为 1 的矩阵。这种现象就是著名的**优惠券收集**（coupon collector）问题实例（Erdos and Renyi 1961）。至于秩的影响情况，可用下面这个结论来说明：对于给定秩为 r 的 $p \times p$ 矩阵，需要大约 $\mathcal{O}(rp)$ 个参数才能精确恢复，因为有 $\mathcal{O}(r)$ 个奇异向量，每个向量的维度为 p。从前面的分析可以看到，对矩阵的**相关性**（coherence）作出一定限制后，基于原子范数松弛方法可准确恢复该矩阵，所需要的样本大小只是维度乘以一个参数因子。相关性可度量矩阵的奇异向量与标准基对齐的程度。

为了理解为什么需要相关性约束，这里通过一个秩为 1 的矩阵 $\boldsymbol{Z} = \boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}$ 来进行解释，该矩阵只在左上角有一个 1，如左式

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{Z}' = \begin{pmatrix} v_1 & v_2 & v_3 & v_4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.16}$$

如果只允许均匀随机地抽取该矩阵的 $N \ll p^2$ 个值，则观测到元素为 0 的可能性很高，进而推断这个矩阵的元素全为 0。类似的情况也会出现在 $\boldsymbol{Z}' = \boldsymbol{e}_1\boldsymbol{v}^{\mathrm{T}}$ 中，其中 $\boldsymbol{v} \in \mathbb{R}^p$ 是任意的 p 维向量，见式（7.16）右式。因此，基于原子范数正则化的任何理论保证必须考虑这些极端情况。矩阵 $\boldsymbol{Z}$ 和 $\boldsymbol{Z}'$ 与 $\mathbb{R}^4$ 中的标准基有最大相关性，这表明它们的左奇异向量和右奇异向量的一些子集与一些标准基向量 $\boldsymbol{e}_j$ 完全对齐。

排除这种特殊矩阵的一种方法是从一些随机合成中抽取矩阵。比如，可构造形如 $\boldsymbol{Z} = \sum_{j=1}^{r} a_j b_j^{\mathrm{T}}$ 的随机矩阵，其中随机向量 $\boldsymbol{a}_j \sim N(0, \boldsymbol{I}_p)$ 和 $\boldsymbol{b}_j \sim N(0, \boldsymbol{I}_p)$

都是独立抽取的。这种随机矩阵的奇异向量极不可能与标准基向量有高度的相关性。Gross（2011）证明，这种合成和采样如果满足不等式

$$N \geqslant Crp\log p \tag{7.17}$$

则用原子范数松弛方法精确恢复矩阵的概率很高，其中 $C > 0$ 为常量。较早但不甚有说服力的证明可参见 Candès and Recht（2009）。更一般地讲，当 C 依赖于奇异向量不相关性时，可能得到精确恢复保证，这种不相关性可像前面那样通过奇异向量和标准基础之间的最大对齐来度量。为进一步了解这类结果，建议读者参考如下文献：Candès and Recht（2009）、Gross（2011）和 Recht（2011），也可参考 Keshavan，Oh and Montanari（2009）对另一种稍微不同的预测所得出的相关结论。

作者进行了一个小型模拟研究，以便更好地理解式（7.17）。设矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的大小为 $p\times r$，它们从独立同分布的标准正太分布中得到元素，并且 $\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$。然后假定缺失值的数量固定不变，并采用 Soft-Impute 算法（有足够小的 λ），使得 $\left\|\mathcal{P}_{\Omega}^{\perp}(\boldsymbol{Z}-\boldsymbol{Z})\right\|_F^2 / \left\|\mathcal{P}_{\Omega}^{\perp}(\boldsymbol{Z})\right\|_F^2 < 10^{-5}$；也就是说，这样可有效恢复观测值。然后检测是否有

$$\left\|\mathcal{P}_{\Omega}^{\perp}(\boldsymbol{Z}-\hat{\boldsymbol{Z}})\right\|_2^2 / \left\|\mathcal{P}_{\Omega}^{\perp}(\boldsymbol{Z})\right\|_2^2 < 10^{-5} \tag{7.18}$$

即对缺失数据插值。将这个过程重复 100 次，每次秩 r 和缺失元素数量不一样。

缺失数据得到成功插值的次数比例如图 7-4 所示。可以看到，当秩比矩阵维数小很多时，有相当高的概率可以恢复缺失元素。但是，当真实的秩过高，恢复会变得很困难。

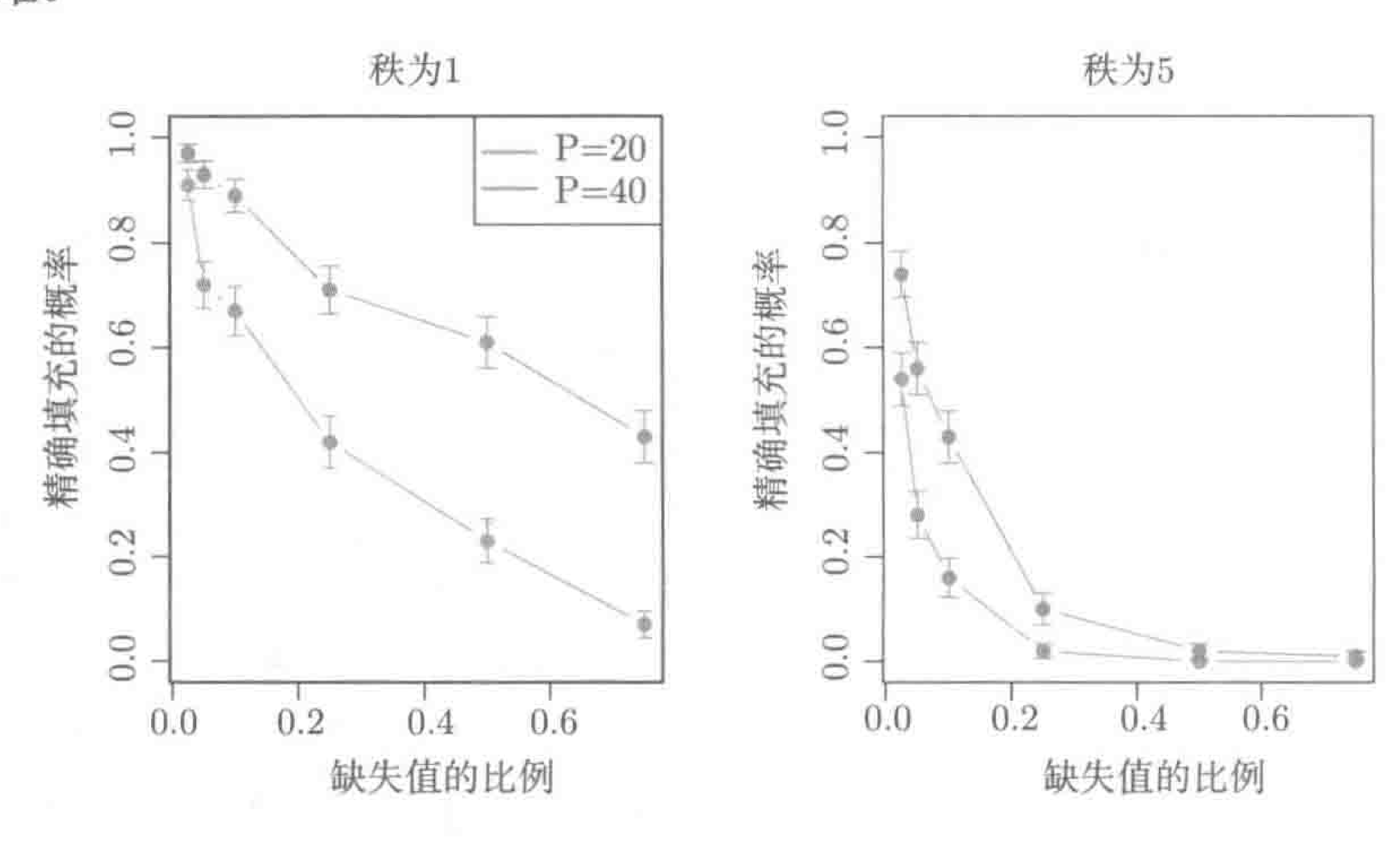


图 7-4　没有噪声时的凸矩阵填充。图中所示为 $n\times n$（$n\in\{20,40\}$）矩阵精确填充的概率（均值 ± 一个标准误差）与缺失值比例之间的函数关系。左图中，所填充矩阵的真实秩为 1，右图为 5

当然，“精确”的情形往往不现实，更合理的假设是观测的元素存在噪声，就如模型（7.7）那样，即 $\boldsymbol{Z} = \boldsymbol{L}^* + \boldsymbol{W}^*$，其中 $\boldsymbol{L}^*$ 的秩为 r。在这种情形下，精确的矩阵填充通常不可能实现，我们的重点在于如何通过式（7.10）得到近似低秩矩阵 $\boldsymbol{L}^*$。奇异向量不相关条件不太适用于有噪声的情形，因为它们对小扰动不具有健壮性。为了理解这一点，可假设矩阵 $\boldsymbol{B}$ 的秩为 $r-1$，其 Frobenius 范数为 1，并有**最大限度的不相关性**，即所有奇异向量与标准基向量正交。对于式（7.16）中的矩阵 $\boldsymbol{Z}$，考虑其扰动矩阵 $\boldsymbol{L}^* = \boldsymbol{B} + \delta\boldsymbol{Z}$（其中 $\delta > 0$）。不管 δ 多小，矩阵 $\boldsymbol{L}^*$ 的秩都为 r，由于标准基向量 $\boldsymbol{e}_1 \in \mathbb{R}^p$ 是 $\boldsymbol{L}^*$ 的一个奇异值向量，所以它就总能保持**最大相关性**。

另一种基于矩阵“尖”比率的目标函数对这些小扰动不敏感（Negahban and Wainwright 2012）。实际上，对任意的非零矩阵 $\boldsymbol{L} \in \mathbb{R}^{p\times p}$，定义 $\alpha_{\mathrm{sp}}(\boldsymbol{L}) = \frac{p\|\boldsymbol{L}\|_\infty}{\|\boldsymbol{L}\|_F}$，其中 $\|\boldsymbol{L}\|_\infty$ 为矩阵 $\boldsymbol{L}$ 中绝对值最大的元素值。这个比率是对矩阵元素取值分散程度的一致度量，其取值范围为 1～ p。当矩阵 $\boldsymbol{L}$ 的所有元素取值都相同时，$\alpha_{\mathrm{sp}}(\boldsymbol{L}) = 1$，这是最小值；而式（7.16）中的矩阵 $\boldsymbol{Z}$ 会得到最大的尖比率 $\alpha_{\mathrm{sp}}(\boldsymbol{L}) = p$。与奇异向量的不相关性相比，尖比率涉及奇异值（以及相应的奇异向量）。因此，只要扰动 $\delta > 0$ 充分小，矩阵 $\boldsymbol{L}^* = \boldsymbol{B} + \delta\boldsymbol{Z}$ 就有较低的尖比率。

对于有尖比率界的原子范数正则化估计子（7.10），Negahban and Wainwright（2012）已证明：所得解 $\hat{\boldsymbol{L}}$ 的上界形如

$$\frac{\left\|\hat{\boldsymbol{L}} - \boldsymbol{L}^*\right\|_F^2}{\|\boldsymbol{L}^*\|_F^2} \leqslant C \ \max\left\{\sigma^2, \alpha_{\mathrm{sp}}^2(\boldsymbol{L}^*)\right\} \frac{rp\log p}{N} \tag{7.19}$$

这个不等式在采样模式和随机噪声上成立的概率很高（假设独立同分布，零均值，并且方差为 σ^2）。另外可参考 Keshavan，Montanarit and Oh（2010）和 Koltchinskii，Lounici and Tsybakov（2011），它们对略有不同的估计子证明了相关的保证。

为了更好地理解式（7.19），作者进行了摸拟实验。定义比率 $\nu = \frac{N}{p^2} \in (0,1)$，表示在 $p\times p$ 矩阵中观测到的元素所占的比例。另外定义秩比率 $\delta = \frac{r\log p}{p}$，这是一个相对的矩阵秩（采用对数因子）。给定固定的噪声方差和尖比率，只要 $\nu > \delta$，则求证式（7.10）所得的结果均方误差（mean-squared error）低。图 7-5 证实了这个结论，并且发现该理论实际上略显保守，详见图题。

图 7-6 给出在噪声情形下矩阵填充纠正误差（imputation error）的另一个例子。采用 Soft-Impute 算法对 40×40 的矩阵求解，这些矩阵是由标准高斯矩阵（其秩分别为 1 和 5）加上标准偏差 $\sigma = 0.5$ 的噪声生成的。对于秩为 1 的情形，即使缺失比例高达 50%，也可恢复缺失值，这些恢复值的平均误差会接近 σ。然而当秩为 5 时，缺失比例 30%，这个算法就会失败。

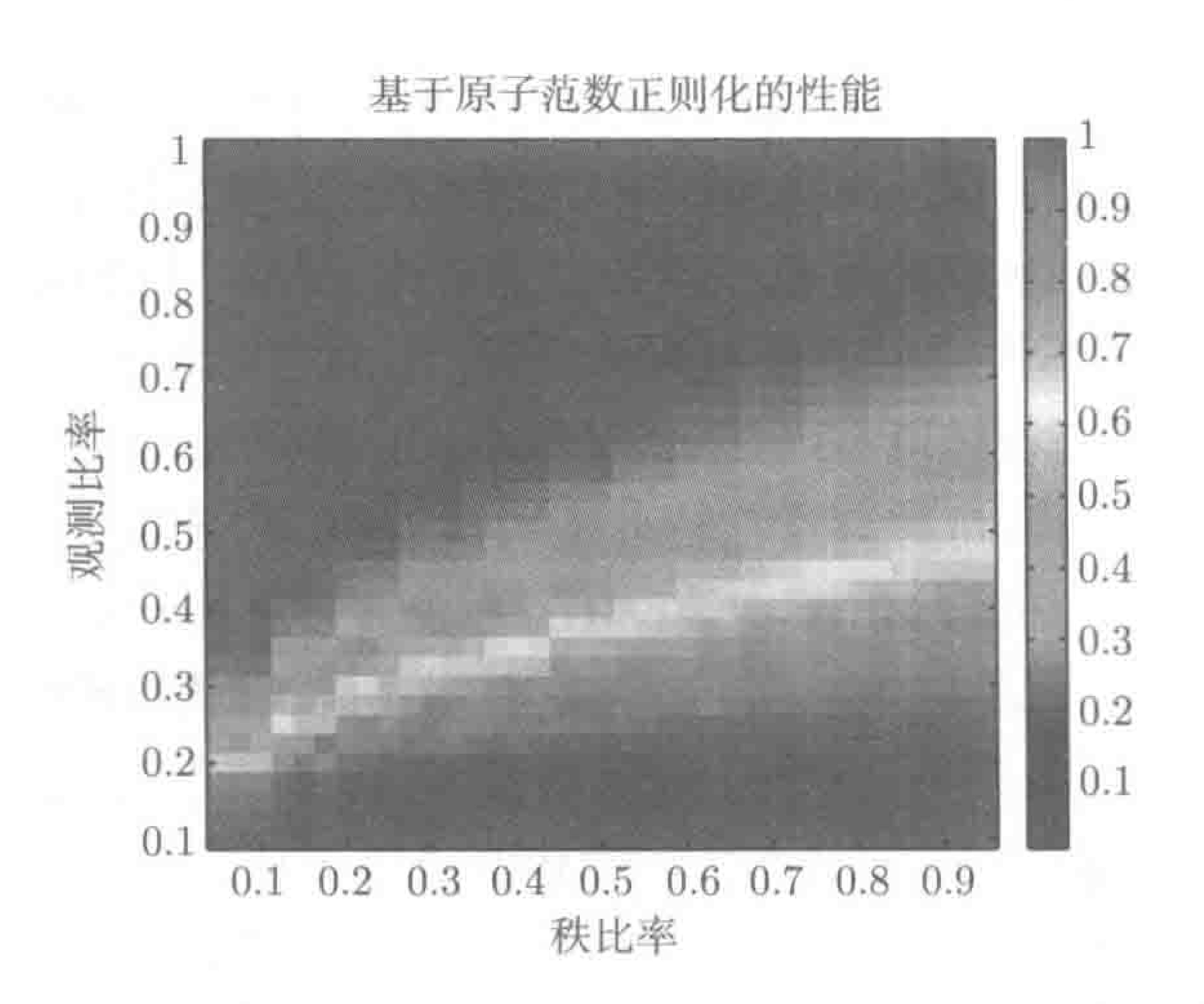


图 7-5　通过 Soft-Impute 算法求解基于原子范数正则化估计子（7.10）的性能。这是一个基于式（7.7）的噪声矩阵填充模型，其中 $\boldsymbol{L}^* = \boldsymbol{C}\boldsymbol{G}^{\mathrm{T}}$ 的秩为 r。这幅图为秩比率 $\delta = \frac{r \log p}{p}$ 和观测比率 $\nu = \frac{N}{p^2}$（该比率表示在 $p \times p$ 矩阵中，观测到的元素所占的比例）的函数曲线图，其中 $p = 50$，采用相对 Frobenius 范数误差 $\frac{\|\hat{\boldsymbol{L}}-\boldsymbol{L}^*\|_F^2}{\|\boldsymbol{L}^*\|_F^2}$ 进行度量。式（7.7）是观测结果的线性形式，其中 $w_{ij} \sim N(0, \sigma^2)$，$\sigma = 1/4$，采用 Soft-Impute 算法求解式（7.10），其中 $\lambda/N = 2\sigma\sqrt{\frac{P}{N}}$，理论上建议选择后者。该理论还指出，只要 $\nu \succ \delta$，Frobenius 误差就会变小，这幅图证实了这个结论（见彩插）

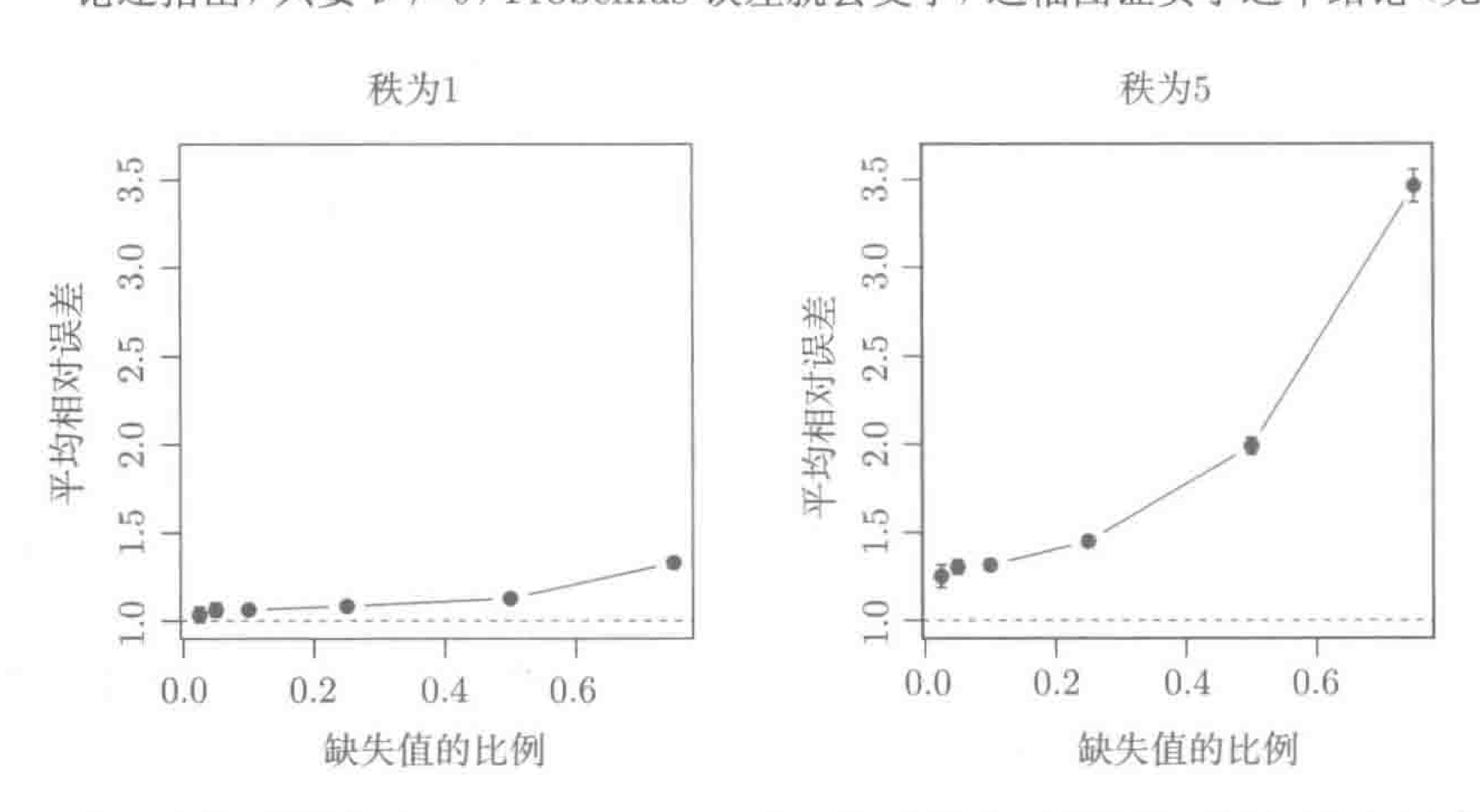


图 7-6　在噪声情形下通过 Soft-Impute 算法实现矩阵填充。该图显示矩阵填充的纠正误差与缺失比例之间的函数关系，其中矩阵大小为 40×40。图中为 100 次模拟的平均绝对误差（$\pm$ 一个标准误差），均相对于噪声标准偏差。为了尽量减少纠正误差，对每种情况都要选择征罚参数。如果通过交叉验证选择参数，所得结果会稍差。左图所填充矩阵的真实秩为 1，右图为 5。左图中，矩阵元素的平均绝对值为 0.80，右图为 1.77

7.3.4 最大间隔分解及相关方法

本节讨论一类与上一节相似的算法。该算法称为**最大间隔矩阵分解**（Maximum Margin Matrix Factorization，MMMF）方法，使用分解模型来逼近矩阵 $\boldsymbol{Z}$（Rennie and Srebro 2005）[①]。矩阵分解的形式为 $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}$，其中矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的大小分别为 $m \times r$ 和 $n \times r$。可通过求解优化问题

$$\underset{\substack{\boldsymbol{A}\in\mathbb{R}^{m\times r}\\ \boldsymbol{B}\in\mathbb{R}^{n\times r}}}{\text{minimize}} \left\{ \left\| \mathcal{P}_{\Omega}(\boldsymbol{Z}) - \mathcal{P}_{\Omega}(\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}) \right\|_F^2 + \lambda \left(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2 \right) \right\} \tag{7.20}$$

来估计这种分解。有趣的是，在 r 足够大的情况下，可证明这个问题与基于原子范数正则化的问题（7.10）等价，在此得出精确解。首先，对于任何矩阵 $\boldsymbol{M}$，可证明（Rennie and Srebro 2005，Mazumder et al. 2010）

$$\|\boldsymbol{M}\|_{\star} = \min_{\substack{\boldsymbol{A}\in\mathbb{R}^{m\times r},\boldsymbol{B}\in\mathbb{R}^{n\times r}\\ \boldsymbol{M}=\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}}} \frac{1}{2}\left(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2 \right) \tag{7.21}$$

习题 7.6 会证明问题（7.21）的解并不唯一。但从式（7.21）可知：对于 $r \geqslant \min(m, n)$，双凸问题（7.20）的解 $\hat{\boldsymbol{M}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{B}}^{\mathrm{T}}$ 与凸问题（7.10）的解一样。更具体地讲，有结论：

定理 1 $\boldsymbol{Z}$ 为 $m \times n$ 矩阵，观测到的元素索引为 Ω。

(a) 若 $r = \min(m, n)$，则式（7.10）的解与基于原子范数正则化的问题（7.10）的解一致（对所有 λ）。

(b) 给定 $\lambda^* > 0$，若式（7.10）的解的秩为 r^*，对于 $r > r^*$，$\lambda = \lambda^*$，式（7.20）有任意最优解为 $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$，则矩阵 $\hat{\boldsymbol{M}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{B}}^{\mathrm{T}}$ 为式（7.10）的最优解。因此，式（7.10）的解空间包含在式（7.20）的解中。

MMMF 算法（7.20）对应的模型有两个参数 (r, λ)，而 Soft-Impute 算法（7.10）对应的模型只有一个参数 λ，由定理 1 可知，该模型族由二维解 $(\hat{\boldsymbol{A}}_{(r,\lambda)}, \hat{\boldsymbol{B}}_{(r,\lambda)})$ 网格上的一条特殊路径形成，图 7-7 描绘了这种情形。图中红点上方的参数组合所得到的任何 MMMF 模型都是多余的，因为与红点处的模型一样。然而在实际应用中，MMMF 并不知道红点在哪，也不知道实际的秩是多少。可通过 $\hat{\boldsymbol{A}}$ 和 $\hat{\boldsymbol{B}}$ 正交得到秩的大小，这仅是一种近似（这与 MMMF 算法的收敛标准有关）。总之，式（7.10）有两个优势：它是凸的，在正则化的同时会尽量减少秩。求解问题（7.20）时，需要选择近似的秩和正则参数 λ。

① “最大间隔”是指基于间隔的损失函数，该函数由这些作者特别定义。虽然本书采用平方误差损失函数，但重点是惩罚项，所以这里会用同样的缩写。

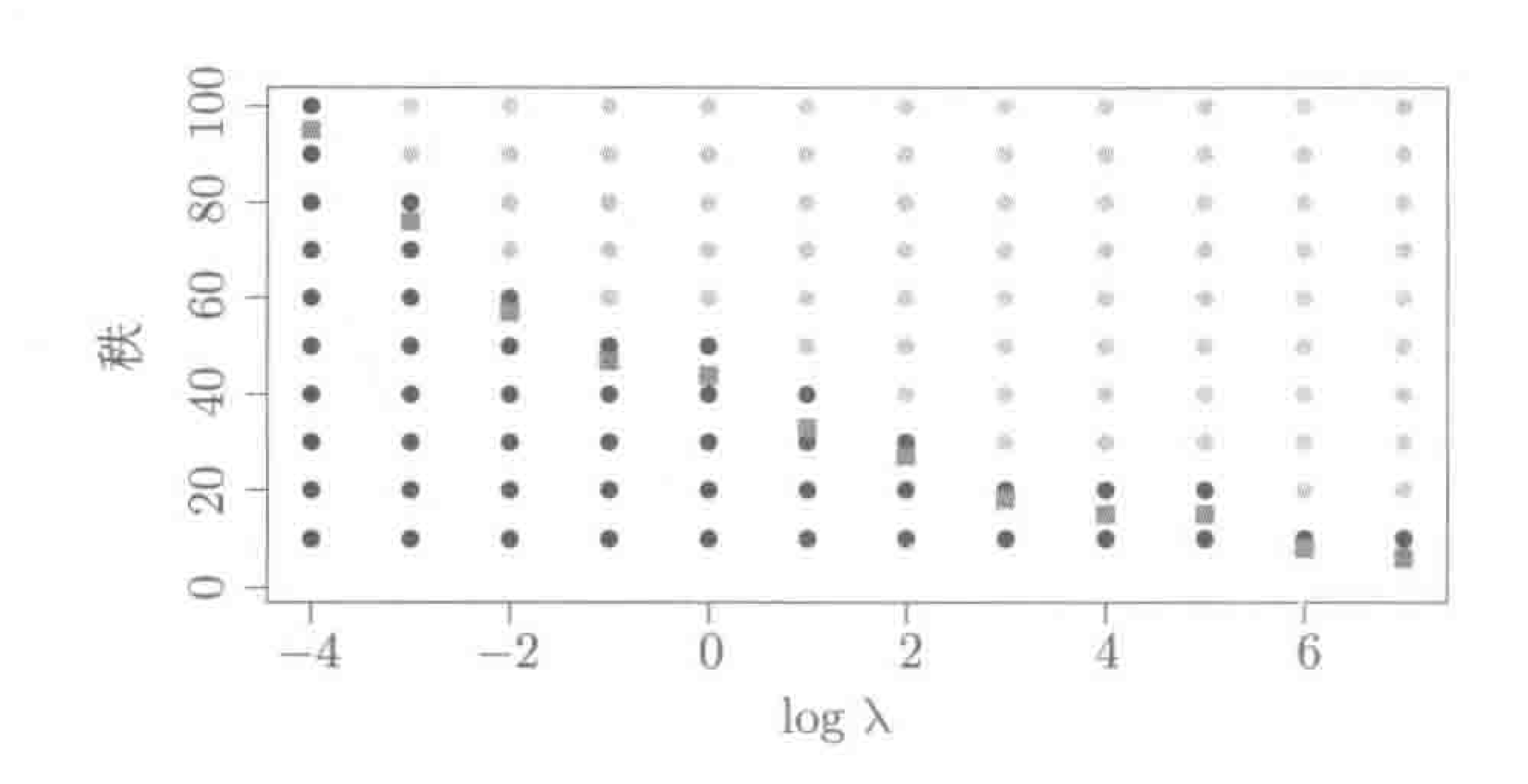


图 7-7 针对一个简单的例子比较 MMMF（灰色和黑色的点）和 Soft-Impute（红色的点）。对于红点上方的秩，MMMF 的解与 Soft-Impute 相同，因此灰点显出了冗余。另一方面，若对 MMMF 固定秩（λ 已指定），且这个秩要比 Soft-Impute 的解小，就会得到一个非凸问题（见彩插）

Keshavan et al.（2010）提出了另一个与之相关的方法，即

$$\left\|\mathcal{P}_\Omega(\boldsymbol{Z})-\mathcal{P}_\Omega(\boldsymbol{USV}^{\mathrm{T}})\right\|_F^2+\lambda\|\boldsymbol{S}\|_F^2 \tag{7.22}$$

需要最小化三元组 $(\boldsymbol{U},\boldsymbol{V},\boldsymbol{S})$，其中 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}_r$，矩阵 $\boldsymbol{S}$ 的大小为 $r\times r$。给定秩 r，可用梯度下降法来最小化式（7.22）。式（7.22）与式（7.20）很像，只不过 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 正交，即所有的“信号”和相应的正则化都包含在（全）矩阵 $\boldsymbol{S}$ 中。与 MMMF 一样，该问题是非凸的，因此使用梯度下降法求解不能保证收敛到全局最优解。此外，它还要针对不同秩进行单独求解。

对于噪声情形下的矩阵填充，Keshavan et al.（2010）针对式（7.22）提出了一些渐近理论，使用缩放会通过宽高比（aspect ratio）m/n 收敛到某个常数 $\alpha\in(0,1)$。下面对其中的一个结果进行简要介绍。设矩阵 $\boldsymbol{Z}$ 的大小为 $m\times n$，该矩阵可写成 $\boldsymbol{Z}=\boldsymbol{U\Sigma V}+\boldsymbol{W}$，其中 $\boldsymbol{\Sigma}\in\mathbb{R}^{r\times r}$ 是对角矩阵。$\boldsymbol{W}$ 是随机矩阵，其元素独立同分布，这些分布的均值为零，方差为 $\sigma^2\sqrt{mn}$。假定矩阵 $\boldsymbol{Z}$ 中的元素是按概率 ρ 独立观测得到。$\hat{\boldsymbol{Z}}$ 是式（7.22）的解，其中 λ 的值是通过优化确定的。Keshavan et al.（2010）已证明：当 $m/n\to\alpha\in(0,1)$ 时，相对误差 $\frac{\|\hat{\boldsymbol{Z}}-\boldsymbol{Z}\|_F^2}{\|\boldsymbol{Z}\|_F^2}$ 会以概率 $1-c(\rho)$ 收敛。如果 $\sigma^2/\rho\geqslant\max_{jj}\Sigma_{jj}$ 则 $c(\rho)$ 为 0，否则就不为 0。这说明估计子发生了相变：如果缺失项的噪声和概率相对于信号强度较低，那么缺失项可以成功恢复。否则，它们在重建缺失项中基本上是无用的。详见 Keshavan et al.（2009）和 Keshavan et al.（2010）。

7.4 减 秩 回 归

本节简要回顾一下 4.3 节的**多元回归**。输出变量为 $y_i\in\mathbb{R}^K$，协变量为 $x_i\in\mathbb{R}^p$，

需要建立 K 个线性回归模型。在 (y_i, x_i) 上有 N 个观测，可用矩阵

$$\boldsymbol{Y} = \boldsymbol{X\Theta} + \boldsymbol{E} \tag{7.23}$$

来表示这一系列的线性回归模型。其中，$\boldsymbol{Y} \in \mathbb{R}^{N\times K}, \boldsymbol{X} \in \mathbb{R}^{N\times p}, \boldsymbol{\Theta} \in \mathbb{R}^{p\times K}$ 为系数矩阵，$\boldsymbol{E} \in \mathbb{R}^{N\times K}$ 是误差矩阵。

最简单的方法是通过 lasso 或弹性网来单独拟和 K 个模型。但输出值可能很多都一样，在拟和 K 个模型时，可以利用这些相似性。4.3 节对每个输出使用组 lasso 的同时选择变量，即使用组 lasso 设置矩阵 $\boldsymbol{\Theta}$ 的整行为零。在本节会假设矩阵 $\boldsymbol{\Theta}$ 为低秩。这就是**多任务**（multitask）机器学习背后的思想。由此得到的模型形如

$$\boldsymbol{Y} = \boldsymbol{XAB}^{\mathrm{T}} + \boldsymbol{E} \tag{7.24}$$

其中 $\boldsymbol{A} \in \mathbb{R}^{p\times r}$，$\boldsymbol{B} \in \mathbb{R}^{K\times r}$。可认为，$r < K$ 所获得的特征 $\boldsymbol{Z} = \boldsymbol{X}\hat{\boldsymbol{A}}$，可通过 K 个单独回归 $\hat{\boldsymbol{Y}} = \boldsymbol{Z}\hat{\boldsymbol{B}}^{\mathrm{T}}$ 在输出中间传递。虽然通过最小二乘拟和是一个非凸优化问题，但通过典型相关分析可得到 $N > p$ 的闭解形式（Hastie et al. 2009）。

例子 7.1 本例关注视频去噪问题。图 7-8 显示了直升机飞越沙漠所拍摄的四幅具有代表性的视频图像。矩阵 $\boldsymbol{Y}$ 的 j 列表示在时间 k 时的一帧图像。整个矩阵 $\boldsymbol{Y}$

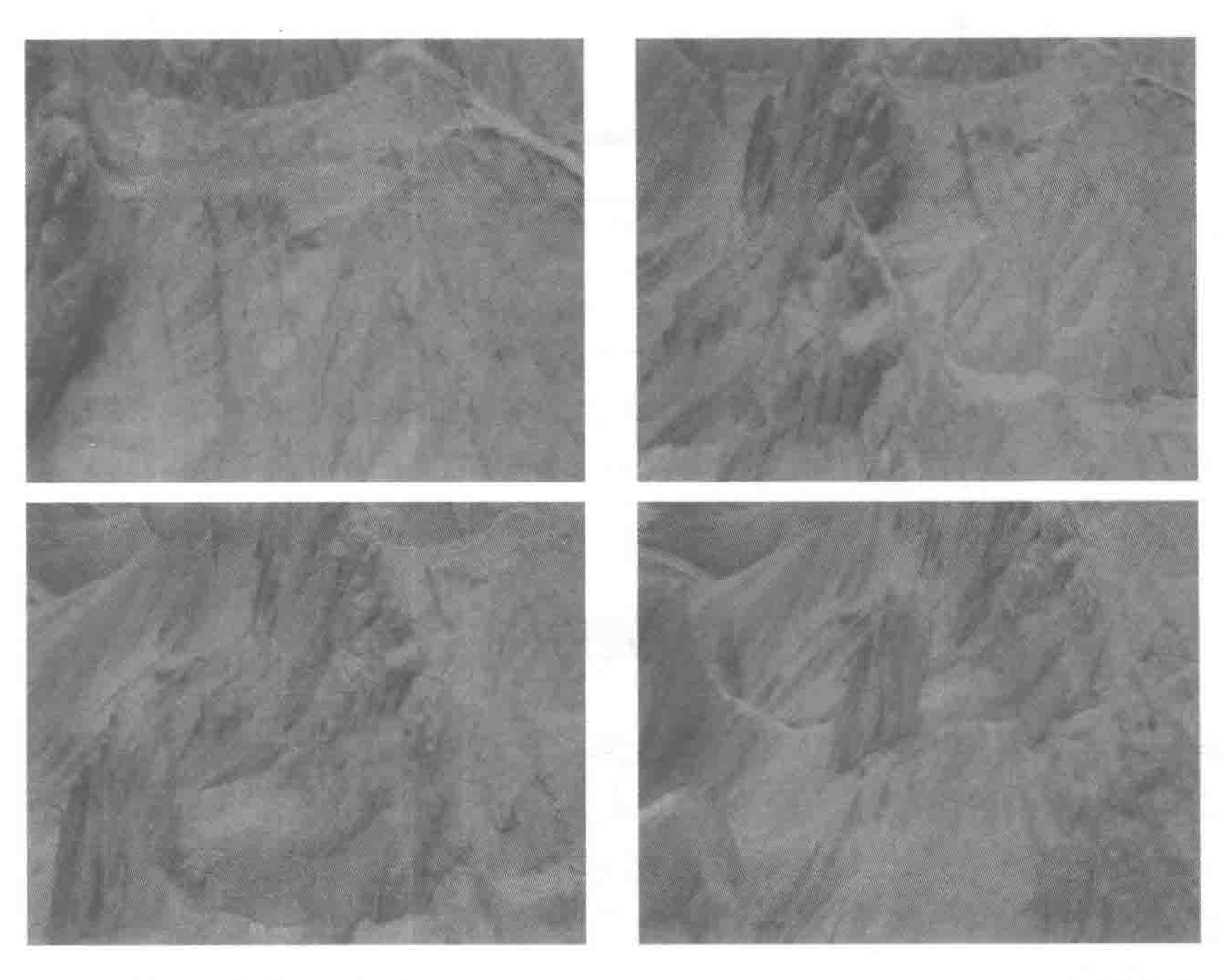


图 7-8 直升机飞越沙漠所拍摄视频中取得的四幅 352×640 图像帧。每幅图像转换为 $\mathbb{R}^N$ 中的一个向量，其中 $N = 352 \times 640 = 225\,280$。一幅图像就是矩阵 $\boldsymbol{Y}$ 中的一列

表示由 K 幅图像帧所构成的视频。矩阵 $\boldsymbol{X}$ 的 p 个列表示图像基函数（见第 10 章中）的字典。当视频序列随时间变化相对缓慢时，矩阵 $\boldsymbol{\Theta}$ 为低秩模型，因此大部分变化可通过少数有代表性图像的线性组合来描述。

图 7-9 给出了用图 7-8 中视频的 $K=100$ 帧图计算的 SVD。虽然矩阵 $\boldsymbol{Y}$ 并不完全是低秩矩阵，但它的奇异值会衰减得很快，这说明通过一个低秩矩阵来对其近似会有很好的效果。

如前所述，在估计时，原子范数可用来作为一个凸惩罚，以得到低秩结构的解。这需要解优化问题

$$\underset{\boldsymbol{\Theta}\in\mathbb{R}^{p\times k}}{\operatorname{minimize}}\left\{\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Theta}\|_F^2+\lambda\|\boldsymbol{\Theta}\|_\star\right\} \tag{7.25}$$

若 λ 充分大，所得的解 $\boldsymbol{\Theta}$ 的秩会小于 $\min(N,K)$。

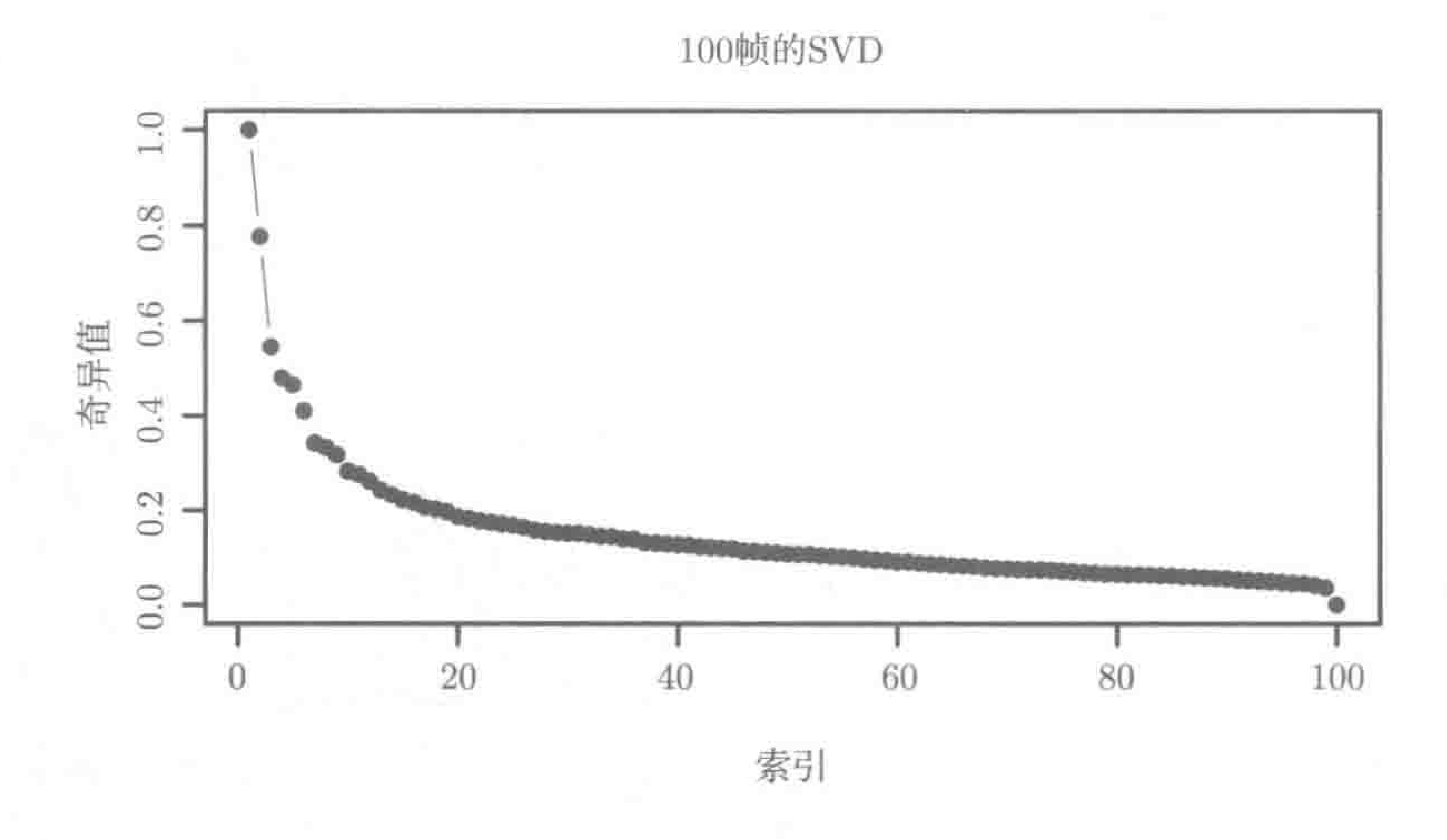


图 7-9　频序列构成的矩阵 $\boldsymbol{Y}\in\mathbb{R}^{p\times K}$（$K=100$ 帧）的奇异值。注意：这些奇异值会迅速衰减，这表明可用低秩矩阵来近似这个矩阵

7.5　通用矩阵回归框架

本节论述通用的“迹”回归框架，矩阵填充和减秩回归都是这个框架的特殊情形。这个通用框架可用统一的理论来处理。

下面从矩阵填充开始介绍该框架。假定 $\boldsymbol{M}$ 为具有部分观测值的 $m\times n$ 矩阵 $\boldsymbol{Z}$ 的模型，我们的目的是填充 $\boldsymbol{Z}$ 中的缺失值。对观测值对 $(\boldsymbol{X}_i,y_i)$，$i=1,2,\ldots,|\Omega|$，可建立模型

$$y_i=\operatorname{trace}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{M})+\varepsilon_i \tag{7.26}$$

其中矩阵 $\boldsymbol{X}_i \in \mathbb{R}^{m\times n}$，$y_i$ 和 ε_i 都是标量。式（7.26）可认为是关于矩阵 $\boldsymbol{X}_i$ 的回归模型，其输出值为 y_i。矩阵的迹内积（trace innner product）相当于向量的内积，其他的概念同普通的回归模型一样 ①。

为了同矩阵填充建立联系，令 $[a(i), b(i)]$ 表示有观测值的第 i 个元素索引。定义 $\boldsymbol{X}_i = e^n_{a(i)} e^{m\mathrm{T}}_{a(i)}$，其中 $e^m_\ell \in \mathbb{R}^m$ 是一个 m 维的单位向量，它的第 ℓ 个元素为 1，其他元素全为 0，即 $\boldsymbol{X}_i$ 除了 $[a(i), b(i)]$ 不为 0，其他元素全为 0。因此，$\mathrm{trace}(\boldsymbol{X}_\mathbf{i}^\mathrm{T}\boldsymbol{M}) = m_{a(i)b(i)}$，即式（7.26）提供了 $\boldsymbol{M}$ 的某些具有噪声 ε_i 的元素（这些元素的索引在 Ω 中）。这里的目的是通过 $\hat{\boldsymbol{M}}$ 来填充 $\boldsymbol{Z}$ 中没有观测到值的元素，这可看成是特征值 $\boldsymbol{X}^*$ 的 $\mathbb{E}(y^*|\boldsymbol{X}^*)$，这些特征值区别于训练集中的其他特征值。

式（7.26）也与更一般的情形相关，因为选择不同的 $\{\boldsymbol{X}_i\}$，可以得到具有低秩约束的不同矩阵估计模型。

上一节介绍的多输出回归模型是另一个例子。输出值和协变量向量通过方程 $y_i = \boldsymbol{\Theta}^\mathrm{T} x_i + \boldsymbol{\varepsilon}_i$（其中，$\boldsymbol{\Theta} \in \mathbb{R}^{p\times K}$ 是回归系数矩阵）联系在一起，$\boldsymbol{\varepsilon}_i \in \mathbb{R}^K$ 为噪声向量。由于每个输出向量 y_i 是由观测值构成的 K 维向量，因而可以用迹的形式重写为 K 个独立观测值的集合。令 $\boldsymbol{X}_{ij} = x_i\left(e^K_j\right)^\mathrm{T}$，其中 $e^K_j \in \mathbb{R}^K$ 是一个单位向量，它的第 j 个元素为 1，则 y_i 的第 j 元素为 $y_{ij} = \mathrm{trace}(\boldsymbol{X}^\mathrm{T}_{ij}\boldsymbol{\Theta}) + \boldsymbol{\varepsilon}_{ij}$。

对于多元回归，矩阵 lasso 形如

$$\underset{\boldsymbol{\Theta}}{\mathrm{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^K \left(y_{ij} - \mathrm{trace}(\boldsymbol{X}^\mathrm{T}_{ij}\boldsymbol{\Theta}) \right)^2 + \lambda \|\boldsymbol{\Theta}\|_\star \right\} \tag{7.27}$$

习题 7.9 给出了另一个例子。可参考 Yuan, Ekici, Lu and Monteiro（2007）、Negahban and Wainwright（2011a）、Rohde Tsybakov（2011），了解更多细节和这种统一框架的好处。另外，Bunea, She and Wegkamp（2011）对减秩多元回归的替代过程进行了分析。

7.6 惩罚矩阵分解

采用最大间隔矩阵分解方法自然会得到其他正则化形式，比如基于 ℓ_1 惩罚版本为

$$\underset{\substack{\boldsymbol{U}\in\mathbb{R}^{m\times r}, \boldsymbol{V}\in\mathbb{R}^{n\times r} \\ \boldsymbol{D}\in\mathbb{R}^{r\times r}}}{\mathrm{minimize}} \left\{ \left\| \boldsymbol{Z} - \mathbf{UDV}^\mathrm{T} \right\|_F^2 + \lambda_1 \|\boldsymbol{U}\|_1 + \lambda_2 \|\boldsymbol{V}\|_1 \right\} \tag{7.28}$$

① 若 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 都是 $m\times n$ 的矩阵，则 $\mathrm{trace}(\boldsymbol{A}^\mathrm{T}\boldsymbol{B}) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}b_{ij}$。

其中 $\boldsymbol{D}$ 为非负对角矩阵。假定 $\boldsymbol{Z}$ 中没有缺失值，可对分解的左奇异向量和右奇异向量采用 ℓ_1 惩罚。这种惩罚的思路是：为了得到可解释性，需要得到奇异向量的稀疏版本。

在讨论如何优化式（7.28）之前，来看看如何使用它。回到 Netflix 的例子，创建一个较小的矩阵，由 1000 用户和 100 部电影构成，每部影片有最高的评分。可使用迭代秩 10 SVD（见 7.3 节）来得到缺失值。将 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的秩设为 2，然后最小化式（7.28），其中 λ_1 和 λ_2 的取值要能让解非常稀疏。所得的解 $\hat{\boldsymbol{V}}$ 有 12 个非零元素，它们的符号相同，表 7-3 为相应的电影。第一组含有浪漫喜剧和动作片，而第二组含有历史动作/奇幻影片。

表 7-3　基于两个惩罚参数的矩阵分解所得到的非零项所对应的电影

第一组	第二组
《婚礼专家》	《指环王：魔戒再现》
《极速 60 秒》	《最后的武士》
《速度与激情》	《指环王：双塔奇兵》
《珍珠港》	《指环王：王者归来》
《曼哈顿灰姑娘》	
《贴身情人》	
《斗气俏冤家》	

如何求解优化问题（7.28）呢？先考虑一维情形，写成如下的约束形式（而不是拉格朗日形式）：

$$\underset{\substack{\boldsymbol{u}\in\mathbb{R}^m,\boldsymbol{v}\in\mathbb{R}^n\\ d\geqslant 0}}{\text{minimize}}\ \left\|\boldsymbol{Z}-d\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right\|_F^2,\quad \text{其约束为 } \|\boldsymbol{u}\|_1\leqslant c_1,\quad \|\boldsymbol{v}\|_1\leqslant c_2 \tag{7.29}$$

其实式（7.29）不是非常有用，因为它往往会产生过于稀疏的解，如图 7-10 右图所示。为了解决这个问题，可增加 ℓ_2 范数作为约束，从而得到最优化问题

$$\begin{aligned}&\underset{\substack{\boldsymbol{u}\in\mathbb{R}^m,\boldsymbol{v}\in\mathbb{R}^n\\ d\geqslant 0}}{\text{minimize}}\ \left\|\boldsymbol{Z}-d\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right\|_F^2\\ &\text{其约束为 } \|\boldsymbol{u}\|_1\leqslant c_1,\ \|\boldsymbol{v}\|_1\leqslant c_2,\ \|\boldsymbol{u}\|_2\leqslant 1,\ \|\boldsymbol{v}\|_2\leqslant 1\end{aligned} \tag{7.30}$$

增加约束可以得到稀疏解，这似乎有点不可思议，但图 7-10 确实验证了这样的观点。

如果我们固定 $\boldsymbol{v}$，则目标函数（7.30）对于 $\boldsymbol{u}$ 是线性的。假设目标函数的线性轮廓呈一定倾斜（如图 7-10 所示），则不完全平行于多面体约束区域的一侧。要求解这个问题，需要将这个线性轮廓尽可能朝右上方移动，同时保证留在约束区域内。解会出现在空心圆上或在灰色轮廓上。注意，若没有 ℓ_2 约束，解会出现在多面体的拐角处，这会让其中的一个系数为 0。从图 7-10 左图可看出，只要 $1\leqslant c_1\leqslant\sqrt{m}$，$1\leqslant c_2\leqslant\sqrt{n}$，这个问题就有明确的定义。

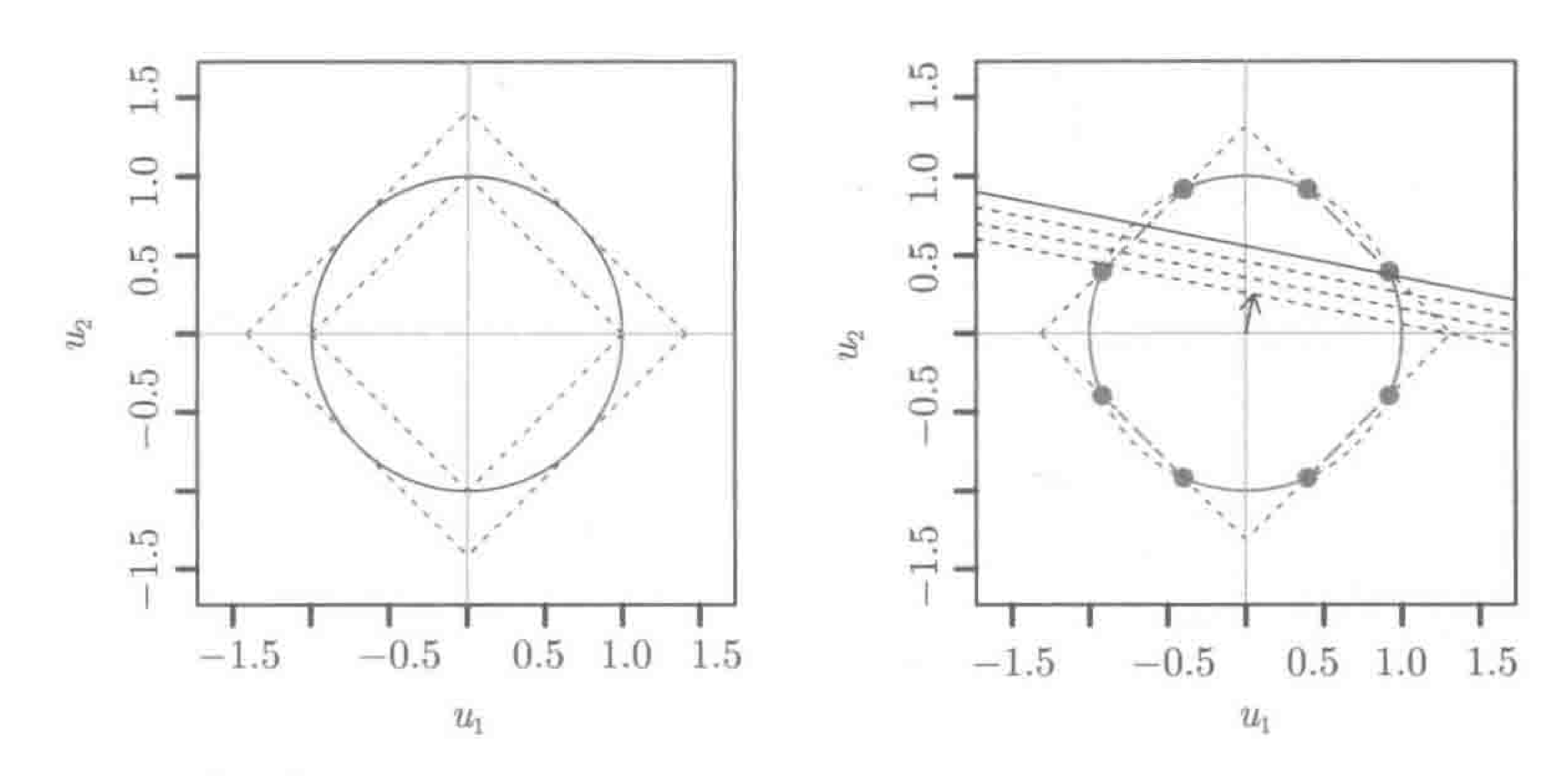


图 7-10 用图形表示 PMD(ℓ_1, ℓ_2) 目标函数下 $\boldsymbol{u} \in \mathbb{R}^2$ 的 ℓ_1 约束和 ℓ_2 约束。这些约束分别为 $\|\boldsymbol{u}\|_2^2 \leqslant 1$ 和 $\|\boldsymbol{u}\|_1 \leqslant c$。一横一纵两条直构成的十字表示坐标轴 u_1 和 u_2。左图的实心圆是 ℓ_2 约束。约束半径必须为 $1 \sim \sqrt{2}$，ℓ_1 约束和 ℓ_2 约束才都会起作用。约束 $\|\boldsymbol{u}\|_1 = 1$ 和 $\|\boldsymbol{u}\|_1 = \sqrt{2}$ 用虚线显示。右图显示 ℓ_1 约束和 ℓ_2 约束，其中 c 为 $1 \sim \sqrt{2}$ 的某个值。红色轮廓为约束区域的边界。黑线是式 (7.30) 作为 $\boldsymbol{u}$ 的函数的线性轮廓，可将其看成在朝右上方移动。红色实线所构成的弧表示在算法 7.2 中 $\lambda_1 = 0$ 时的解（ℓ_2 约束起作用，而 ℓ_1 约束没有起作用）。该图展示的是二维情形，对于 ℓ_1 约束和 ℓ_2 约束都会起作用的点，它们的坐标 u_1 和 u_2 都不会为 0。没有 ℓ_2 约束，结束总会在拐角处，这样会得到平凡解（见彩插）

由于式 (7.30) 具有半凸性，可用交替迭代方式来求解它。很容易验证，在每个方向上的解是一个软阈值操作。例如，对于 $\boldsymbol{v} \in \mathbb{R}^n$，其更新形式为

$$\boldsymbol{u} \leftarrow \frac{S_{\lambda_1}(\boldsymbol{Z}\boldsymbol{v})}{\|S_{\lambda_1}(\boldsymbol{Z}\boldsymbol{v})\|_2} \tag{7.31}$$

对它的向量参数可逐元素地采用软阈值算子 $\mathcal{S}$。式 (7.31) 中所选择的阈值 λ_1 必须满足这样的约束：在 $\|\boldsymbol{u}\|_1 \leqslant c_1$ 中，它会设置为 0；否则要让 λ_1 为一个正常数，使 $\|\boldsymbol{u}\|_1 = c_1$（见习题 7.7）。整个过程见算法 7.2。注意，如果 $c_1 > \sqrt{m}$ 且 $c_2 > \sqrt{n}$，会使 ℓ_1 约束无效，则算法 7.2 会退化成用 power 方法计算矩阵 $\boldsymbol{Z}$ 的最大奇异向量。5.9 节详细介绍了 power 方法。对于迭代软阈值更新（与算法 7.2 相关），最近的一些工作已经给出了理论保证，参见参考文献注释。

目标函数 (7.30) 非常有用，可对 $\boldsymbol{u}$ 或 $\boldsymbol{v}$ 当中的任意一个采用其他惩罚（除 ℓ_1 范数以外），比如融合 lasso 惩罚

$$\Phi(\boldsymbol{u}) = \sum_{j=2}^{m} |u_j - u_{j-1}| \tag{7.32}$$

其中 $\boldsymbol{u} = (u_1, u_2, \ldots, u_m)$。这种选择对于沿每维依次 ($j = 1, 2, \ldots, m$) 进行强制光滑很实用，例如在基因组学应用中的染色体位置。因为选择了这种惩罚，所以算法

7.2 中的最小化步骤也必须相应地修改。此外，可以修改算法 7.2 来处理矩阵的缺失值，比如计算内积 $\boldsymbol{Z}\boldsymbol{v}$ 和 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{v}$ 时，可以通过删除缺失值来处理缺失项。

算法 7.2　交替软阈值的秩 1 惩罚矩阵分解

1. 设 $\boldsymbol{v}$ 是 $\boldsymbol{Z}$ 的最大奇异值对应的左奇异向量。
2. 执行更新 $\boldsymbol{u} \leftarrow \dfrac{\mathcal{S}_{\lambda_1}(\boldsymbol{Z}\boldsymbol{v})}{\|\mathcal{S}_{\lambda_1}(\boldsymbol{Z}\boldsymbol{v})\|_2}$，其中 λ_1 是使 $\|\boldsymbol{u}\|_1 \leqslant c_1$ 最小的值。
3. 执行更新 $\boldsymbol{v} \leftarrow \dfrac{\mathcal{S}_{\lambda_1}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{u})}{\|\mathcal{S}_{\lambda_1}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{u})\|_2}$，其中 λ_2 是使 $\|\boldsymbol{v}\|_1 \leqslant c_2$ 最小的值。
4. 反复迭代步骤 2 和 3，直到收敛。
5. 返回 $\boldsymbol{u}$ 和 $\boldsymbol{v}$，使得 $d = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{v}$。

为了得到多因子惩罚的矩阵分解，可对矩阵 $\boldsymbol{Z}$ 连续采用秩 1 算法（7.2），如算法 7.3 所示。设算法 7.2 中 $\lambda_1 = \lambda_2 = 0$，则可证明 K 分解 PMD 算法会导致 $\boldsymbol{Z}$ 的 K 秩 SVD。实际上，依次得到的这些解是正交的。随着惩罚的出现，解不再出现在 $\boldsymbol{Z}$ 的列和行的空间中，因此得到的解不再正交。

算法 7.3　多因子惩罚的矩阵分解

1. 设 $\boldsymbol{R} \leftarrow \boldsymbol{Z}$
2. $k = 1, \ldots, K$
   (a) 对数据 $\boldsymbol{R}$，采用算法 7.2 中的单个分解方法来得到 $\boldsymbol{u}_k$、$\boldsymbol{v}_k$ 和 $\boldsymbol{d}_k$。
   (b) 更新 $\boldsymbol{R} \leftarrow \boldsymbol{R} - d_k\boldsymbol{u}_k\boldsymbol{v}_k^{\mathrm{T}}$

注意，前面讨论的稀疏矩阵分解和矩阵填充之间的区别很重要。为了成功填充矩阵，需要 $\boldsymbol{Z}$ 的奇异向量具有较低的相关性。也就是说，它们需要**稠密性**。在稀疏矩阵分解中，为了让解具有可解释性，需要得到**稀疏**的奇异向量。这种情形下，矩阵填充并不是主要目标。

不同于凸函数的最小化，交替最小化双凸函数不能保证得到全局最优解。在特殊情况下可证明，只要初始向量不与解正交，算法就会收敛到最优解，例如计算最大奇异向量的 power 方法。但在一般情况下，这些算法只保证能得到函数的局部最优解，5.9 节详细介绍过这个问题。根据经验来看，这些算法在实践中表现得非常好，最近的一些理论工作为这种现象提供了严格的理论保证，可见阅本章最后的参考文献注释一节。

Lee, Shen, Huang and Marron（2010）提出针对两路数据的双聚类（biclustering）采用惩罚矩阵分解。第 8 章会介绍如何应用惩罚矩阵分解得到惩罚多元方法，如主成分的稀疏版本、典型相关性和聚类等。

7.7　矩阵分解的相加形式

在基于加法的矩阵分解问题中，矩阵可分解成两个或更多的矩阵之和。这些相

加的各个部分应具有互补的结构，比如，研究最广泛的一种情形是将矩阵分解成稀疏矩阵与低秩矩阵之和（参见 9.5 节）。矩阵分解的相加形式有各种各样的应用，包括后面所讨论的因子分析、健壮 PCA、矩阵填充和多元回归等。

这些应用中的大多数可以用噪声线性观测模型 $\boldsymbol{Z} = \boldsymbol{L}^* + \boldsymbol{S}^* + \boldsymbol{W}$ 来描述，其中 $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ 分别表示低秩矩阵和稀疏矩阵，$\boldsymbol{W}$ 表示噪声矩阵。在某些情形下，可以考虑这个模型的广义形式，即 $\mathfrak{X}(\boldsymbol{L}^* + \boldsymbol{S}^*)$ 的噪声版本，其中，$\mathfrak{X}$ 是相加矩阵上的某种线性算子（比如，矩阵填充中的投影算子 $\mathcal{P}_{\Omega}$，或通过矩阵回归中的模型矩阵 $\boldsymbol{X}$ 进行的乘法）。

通过公式

$$\underset{\substack{\boldsymbol{L} \in \mathbb{R}^{m \times n} \\ \boldsymbol{S} \in \mathbb{R}^{m \times n}}}{\operatorname{minimize}} \left\{ \frac{1}{2} \|\boldsymbol{Z} - (\boldsymbol{L} + \boldsymbol{S})\|_F^2 + \lambda_1 \Phi_1(\boldsymbol{L}) + \lambda_2 \Phi_2(\boldsymbol{S}) \right\} \tag{7.33}$$

可以估计 $(\boldsymbol{L}^*, \boldsymbol{S}^*)$，其中 Φ_1 和 Φ_2 是惩罚函数，用于实现不同的广义稀疏。例如，在有低秩矩阵和稀疏矩阵的情形中，可分别令 $\Phi_1(\boldsymbol{L}) = \|\boldsymbol{L}\|_*$，$\Phi_2(\boldsymbol{S}) = \|\boldsymbol{S}\|_1$。

下面介绍矩阵分解相加形式的一些应用。

有稀疏噪声的因子分析。因子分析是一种广泛使用的线性降维形式，线性降维是广义的主成分分析。因子分析可以理解为这样一个生成模型：使用“噪声子空间”模型

$$y_i = \mu + \boldsymbol{\Gamma} u_i + w_i, \quad i = 1, 2, ..., N \tag{7.34}$$

生成随机向量 $y_i \in \mathbb{R}^p$，其中 $\mu \in \mathbb{R}^p$ 为均值向量，$\boldsymbol{\Gamma} \in \mathbb{R}^{p \times r}$ 为载荷矩阵（loading matrix）。随机变量 $u_i \sim N(0, I_{r \times r})$ 和 $w_i \sim N(0, \boldsymbol{S}^*)$ 是独立的。由 $\boldsymbol{\Gamma}$ 的列张成的 r 维子空间会生成一个随机元素来得到模型（7.34）的每个向量 y_i。给定来自该模型的 N 个样本，目标是估计载荷矩阵 $\boldsymbol{\Gamma}$ 的列，即秩为 r 的矩阵 $\boldsymbol{L}^* = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} \in \mathbb{R}^{p \times p}$ 张成 $\boldsymbol{\Gamma}$ 列空间。

可证明 y_i 的协方差矩阵为 $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} + \boldsymbol{S}^*$。因此，若在 $\boldsymbol{S}^* = \sigma^2 \boldsymbol{I}_{p \times p}$ 的特殊情形下，则 $\boldsymbol{\Gamma}$ 的列张成的空间等价于 $\boldsymbol{\Sigma}$ 的前 r 个特征向量张成的空间。因此，通过标准的主成分分析来得到 $\boldsymbol{\Gamma}$。实际上，通过计算数据矩阵 $\boldsymbol{Y} \in \mathbb{R}^{N \times p}$ 的 SVD 可实现主成分析，这在 7.2 节中介绍过。$\boldsymbol{Y}$ 的右奇异向量为样本协方差矩阵的特征向量，这是对 $\boldsymbol{\Sigma}$ 的一致估计。

如果协方差矩阵 $\boldsymbol{S}^*$ 不等于一个数乘以单位矩阵，又该如何？通常在因子分析中会假定 $\boldsymbol{S}^*$ 是对角矩阵，但噪声方差取决于该数据的成分。一般而言，可能有非零的非对角线元素，但数量较少，所以可以认为它是稀疏矩阵。在这种情形下，不能保证 $\boldsymbol{\Gamma}$ 的列张成的空间等价于 $\boldsymbol{\Sigma}$ 的前 r 个特征向量张成的空间。若不是这种情况，则 PCA 会**不一致**，这意味着它无法得到真正的列张成的空间，即使有无限的样本数量。

如果 $\boldsymbol{S}^*$ 是方阵，那么要估计的问题 $\boldsymbol{L}^* = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}$ 可以看成是一般观察模型在 $p = N$ 时的实例。在给定 $\{y_i\}_{i=1}^N$ 时，观测矩阵 $\boldsymbol{Z} \in \mathbb{R}^{p\times p}$ 为样本协方差矩阵 $\frac{1}{N}\sum_{i=1}^N y_i y_i^{\mathrm{T}}$。因此有 $\boldsymbol{Z} = \boldsymbol{L}^* + \boldsymbol{S}^* + \boldsymbol{W}$，其中 $\boldsymbol{L}^* = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}$ 的秩为 r，随机矩阵 $\boldsymbol{W}$ 是 Wishart 噪声的重新中心化，即零均值矩阵 $\boldsymbol{W} = \frac{1}{N}\sum_{i=1}^N y_i y_i^{\mathrm{T}} - (\boldsymbol{L}^* + \boldsymbol{S}^*)$。

健壮 PCA。在 7.2 节介绍过，标准主成分分析可通过数据矩阵 $\boldsymbol{Z} \in \mathbb{R}^{N\times p}$（列中心化）的 SVD 来实现，其中第 i 行表示第 i 个样本（每个样本是 p 维向量）。由前面方程可知，最小化 $\|\boldsymbol{Z} - \boldsymbol{L}\|_F^2$（将 $\boldsymbol{L}$ 的秩作为约束）可以得到秩为 r 的矩阵 $\boldsymbol{Z}$ 的 SVD 分解。如果数据矩阵 $\boldsymbol{Z}$ 的某些元素被损坏怎么办？或者出现更糟的情形：数据矩阵行（数据向量）的某个子集被损坏怎么办？PCA 的目标函数是二次函数，它的解（秩 r 的 SVD）对这些扰动很敏感。

矩阵分解的相加形式是得到健壮 PCA 的一种方式。具体而言，这里不是用一个低秩矩阵近似 $\boldsymbol{Z}$，而是用一个具有稀疏成分的低秩矩阵（该矩阵可由 $\boldsymbol{L} + \boldsymbol{S}$ 得到）对损坏的数据建模。在某些元素被损坏的情况下，可通过逐元素稀疏对分量 $\boldsymbol{S}$ 建模，使其具有相对少的非零项；而逐行被损坏的情况更有挑战性，可用行稀疏矩阵来建模。给定秩 r 和稀疏性 k，可直接求解优化问题

$$\underset{\substack{\mathrm{rank}(\boldsymbol{L})\leqslant r \\ \mathrm{card}(\boldsymbol{S})\leqslant k}}{\mathrm{minimize}} \frac{1}{2}\|\boldsymbol{Z} - (\boldsymbol{L} + \boldsymbol{S})\|_F^2 \tag{7.35}$$

这里的 card 表示基数约束，或是非零项的总数（在逐元素被损坏的情况），或是非零行的总数（在逐行被损坏的情况）。当然，目标函数（7.35）的约束分别为矩阵的秩约束和基数约束，因此该目标函数属于双非凸问题。可分别令式（7.33）中的 $\Phi_1(\boldsymbol{L}) = \|\boldsymbol{L}\|_\star$ 和 $\Phi_2(\boldsymbol{S}) = \sum_{i,j}|s_{ij}|$，自然得到式（7.35）的一种凸松弛形式。

图 7-11 给出了一个具有上述惩罚的健壮 PCA 的例子。这些内容取自 Mazumder 和 Hastie 没有发表的论文（图像来自 Li, Huang, Gu and Tian 2004）。数据矩阵 $\boldsymbol{Z}$ 的列取自视频监控画面的帧，有噪声，并有缺失的像素值（接下来会介绍这方面的内容）。这幅图的最后两列为重构的帧，低秩部分表示静态背景，而稀疏部分是每帧的变化情况（在这里就是人的移动）。

健壮矩阵填充。健壮性也是矩阵填充（见 7.3 节）所关心的问题，矩阵填充在协同过滤和推荐系统中经常使用。评分被损坏的原因有很多，比如用户可能会故意滥用系统（例如，一个电影明星想让自己的电影在 Netflix 上得到更多推荐）。另外，一部分用户可能在系统上捣乱，比如 2002 年《纽约时报》报道，亚马逊的系统被对手操纵，向对基督教文化感兴趣的用户推荐性爱手册（Olsen 2002）。

具有健壮性的矩阵填充可以基于健壮 PCA 的原理，即在表示模型时引入稀疏部分 $\boldsymbol{S}$。稀疏的性质取决于建模要求：如果认为只有一小部分元素被损坏，则通过 ℓ_1 范数得到逐元素稀疏性即可；如果要考虑对用户（行）建模，则要考虑逐行稀疏

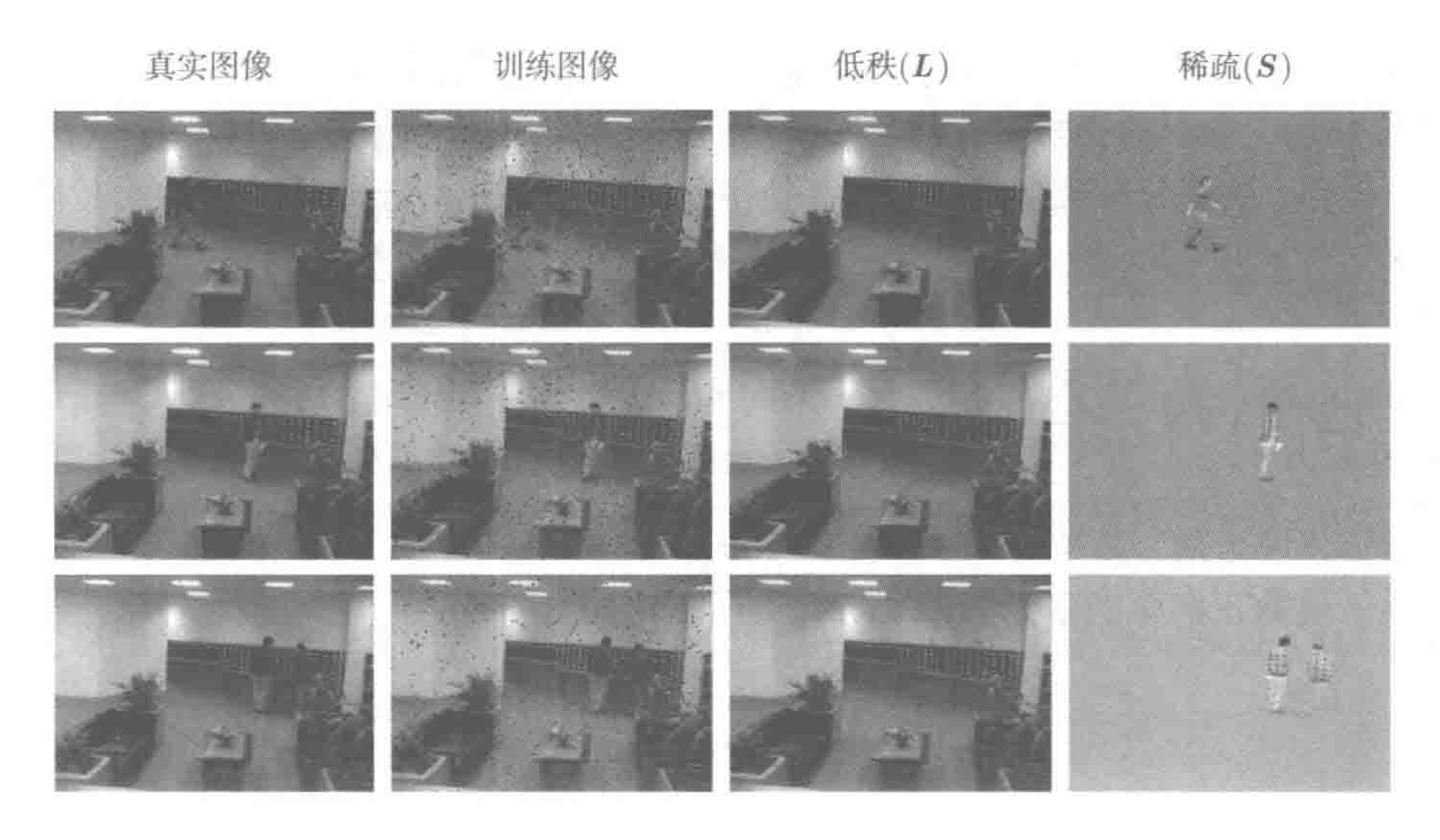


图 7-11 视频监控。该图分别展示了真实形象、有缺失值和噪声的训练图像、被估计的低秩部分，以及一起对齐的稀疏部分。真实图像从视频中采样得到，包括具有不同光照和基准测试一系列图像。尽管图像有缺失部分，并且这个过程还有噪声，但运动部分（人）还是能从固定背景中成功分离出来

性，比如组 lasso 范数 $\|\boldsymbol{S}\|_{1,2}=\sum_{i=1}^{m}\|\boldsymbol{S}_i\|_2$，其中 $\boldsymbol{S}_i\in\mathbb{R}^n$ 表示矩阵的第 i 行。这种选择会将式（7.10）修改成

$$\underset{\boldsymbol{L},\boldsymbol{S}\in\mathbb{R}^{m\times n}}{\text{minimize}}\left\{\frac{1}{2}\sum_{(i,j)\in\Omega}\left(z_{ij}-(\boldsymbol{L}_{ij}+\boldsymbol{S}_{ij})\right)^2+\lambda_1\|\boldsymbol{L}\|_\star+\lambda_2\sum_{i=1}^{m}\|\boldsymbol{S}_i\|_2\right\} \tag{7.36}$$

习题 7.10 会证明这个公式与原子范数正则化的健壮 Huber 损失函数等价。基于 7.1 节的 Soft-Impute 可以得到一种用 Huber 损失函数替代平方和损失函数的算法。图 7-11 显示了该方法应用于视频监控的结果数据。

多元回归。多元线性回归模型为 $y_i=\boldsymbol{\Theta}^{\mathrm{T}}x_i+\varepsilon_i$，其中 $\boldsymbol{\Theta}\in\mathbb{R}^{p\times K}$ 是用来预测多元输出向量 $y\in\mathbb{R}^K$ 的回归系数矩阵。7.5 介绍过回归矩阵的一个应用。输出矩阵 $\boldsymbol{Y}$ 的每一列表示一个经过向量化的图像，即整个矩阵表示有 K 帧的视频序列。模型矩阵 $\boldsymbol{X}$ 表示有 p 个图像的基函数，一列一个，比如每个列表示在不同尺度下、不同位置上的二维小波正交基（见 10.2.3 节）。如图 7-8 所示，在视频序列中，矩阵 $\boldsymbol{Y}$ 的奇异值会迅速衰减，因此可以用低秩矩阵来进行很好的近似。

在更接近实际的情况下，视频序列由两部分组成：背景和各种类型的前景元素。背景往往变化缓慢，从而适合用低秩模型来刻画；而前景元素变化更为迅速，并且可能会消失并重新出现。（在图 7-8 中，可认为直升飞机拍摄的视频具有纯背景）。因此，视频序列的更实用的模型应具有的分解形式为 $\boldsymbol{\Theta}=\boldsymbol{L}+\boldsymbol{S}$，其中 $\boldsymbol{L}$ 是

低秩矩阵，$\boldsymbol{S}$ 是相对稀疏的矩阵。可以用 $\boldsymbol{S}$ 的活动元素 [包含基函数（行）和时间位置（列）] 来表示视频中的前景元素。

当然，这些分解类型也出现在其他多元回归应用中。在一般情形下，可使用估计子

$$\underset{\boldsymbol{L},\boldsymbol{S}}{\text{minimize}}\left\{\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{K}\left(y_{ij}-\text{trace}(\boldsymbol{X}_{ij}^{\mathrm{T}}(\boldsymbol{L}+\boldsymbol{S}))\right)^{2}+\lambda_1\|\boldsymbol{L}\|_{\star}+\lambda_2\|\boldsymbol{S}\|_1\right\} \tag{7.37}$$

来尝试恢复分解，其中 $\boldsymbol{X}_{ij}=x_i e_j^{K\mathrm{T}}$（$i=1,\ldots,N$，$j=1,\ldots,k$。注意，这是对基于原子范数正则化的多元回归（7.27）的自然推广。

参考文献注释

Fazel（2002）早期的研究工作将原子范数作为秩约束的替代品。Srebro, Alon and Jaakkola（2005）研究了原子范数，以及矩阵填充、协同过滤情形下的秩约束相关松弛问题。Bach（2008）对原子范数正则化的一致性得出了一些渐近理论。Recht, Fazel and Parrilo（2010）对压缩感知观测模型得出了原子范数松弛性能的非渐近边界。论文 Negahban and Wainwright（2011a）和 Rohde and Tsybakov（2011）对原子范数松弛在更一般的观察模型下进行了非渐近分析。

最大间隔的矩阵分解在这些文献中讨论过：Srebro and Jaakkola（2003）、Srebro, Alon and Jaakkola（2005）、Srebro, Rennie and Jaakkola（2005）。谱正则化和 Soft-Impute 算法由 Mazumder et al.（2010）提出。惩罚矩阵分解在 Witten, Tibshirani and Hastie（2009）中进行了介绍。有多名研究人员研究过基于原子范数的矩阵填充。Srebro, Alon and Jaakkola（2005）给出了预测误差界的最初的结果。对低秩矩阵在无噪声下的精确恢复的第一个理论成果由 Candès and Recht（2009）提出，随后有多名研究人员进行了相关改进。Gross（2011）提出一种双见证（dual-wintness）方案，证明给定的无噪声观察任意基的原子范数松弛，并推广到逐元素采样的情形，见 Recht（2011）的相关讨论。Keshavan et al.（2009）对一个稍有不同的两阶段过程提供了精确的恢复保证，这个过程会修整矩阵的某些行和列，然后再进行 SVD。噪声观测模型更接近现实情形，有很多文献对其展开了研究，比如 Candès and Plan（2010）、Negahban and Wainwright（2012）和 Keshavan et al.（2010）。

无噪声的矩阵分解相加形式最早由 Chandrasekaran, Sanghavi, Parrilo and Willsky（2011）提出，他们得出了最坏情况精确恢复任意低秩/稀疏对的不相关条件。Candès, Li, Ma and Wright（2011）的后续工作研究了稀疏随机扰动的情况下，低秩矩阵和健壮 PCA 的应用。Xu, Caramanis and Sanghavi（2012）提出了另一种

健壮 PCA 方法，对有损坏的行稀疏矩阵建模。Chandrasekaran, Parrilo and Willsky（2012）提出针对隐变量的高斯图模型问题的稀疏/低秩分解方法。对于更一般的噪声环境，Hsu, Kakade and Zhang（2011）和 Agarwal, Negahban and Wainwright（2012b）提供了式（7.33）的相对界。

最近的研究成果拓展了应用于非凸的问题的交替最小化算法理论，涉及的问题包括矩阵填充（Netrapalli，Jain and Sanghvi 2013）、相检索（Netrapalli et al. 2013）、回归混合（Yi, Caramanis and Sanghavi 2014）和字典学习（Agarwal, Anandkumar, Jain, Netrapalli and Tandon 2014）。这些论文证明：给予适当的初始化，交替最小化方法会（以高概率）收敛于有类似于全局最小的统计精度估计。同样，对于稀疏特征向量的恢复，基于软阈值的改进型 power 方法也有理论保证（Ma 2013 和 Yuan and Zhang 2013）。

习　题

习题 7.1　矩阵的奇异值分解见式（7.2）。请解答下列问题。

(a) 求证：对矩阵 $\boldsymbol{Z}$ 按列进行中心化后得到的 SVD 是 $\boldsymbol{Z}$ 的主成分。

(b) 求证：连续 PC 不相关的条件等价于向量 $\{\boldsymbol{v}_j\}$ 正交。7.2 节中向量 $\{\boldsymbol{s}_j\}$ 与 SVD 各个成分之间是什么关系？

习题 7.2　这个习题需要通过式（7.3）的结论来完成证明工作，这个结果为

$$\hat{\boldsymbol{Z}}_r = \underset{\text{rank}(\boldsymbol{M})=r}{\arg\min} \|\boldsymbol{Z}-\boldsymbol{M}\|_F^2$$

其中 $\hat{\boldsymbol{Z}}_r = \boldsymbol{U}\boldsymbol{D}_r\boldsymbol{V}^{\mathrm{T}}$ 取 SVD 前 r 个分量（SVD 的形式为 $\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$，$\boldsymbol{D}_r$ 与 $\boldsymbol{D}$ 一样，只是将 $\boldsymbol{D}$ 中除前 r 个对角线元素之外的所有对角线元素置为 0）。这里假设 $m \leqslant n$，且 $\text{rank}(\boldsymbol{Z}=m)$。

首先要注意：任何秩为 r 的矩阵 $\boldsymbol{M}$ 都可以分解为 $\boldsymbol{M}=\boldsymbol{Q}\boldsymbol{A}$，其中 $\boldsymbol{Q}=\mathbb{R}^{m\times n}$ 是正交矩阵，$\boldsymbol{A}=\mathbb{R}^{r\times n}$。

(a) 求证：给定 $\boldsymbol{Q}$，则 $\boldsymbol{A}$ 的最优值为 $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Z}$。

(b) 基于 (a) 的结论，求证：最小化 $\|\boldsymbol{Z}-\boldsymbol{M}\|_F^2$ 等价于求解

$$\underset{\boldsymbol{Q}\in\mathbb{R}^{m\times r}}{\text{maximize}}\ \text{trace}(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{Q}),\quad 其约束为\ \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}=\boldsymbol{I}_r \tag{7.38}$$

其中 $\boldsymbol{\Sigma}=\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$。

(c) 求证：式（7.38）等价于问题

$$\underset{\boldsymbol{Q}\in\mathbb{R}^{m\times r}}{\text{maximize}}\ \text{trace}(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{Q}),\quad 其约束为\ \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}=\boldsymbol{I}_r \tag{7.39}$$

(d) 给定正交矩阵 $\boldsymbol{Q} = \mathbb{R}^{m\times r}$，令 $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}$，对角元素为 $h_{ii}(i = 1,\dots,m)$。求证若 $h_{ii} \in [0,1]$，$\sum_{i=1}^m h_{ii} = r$，则式（7.39）与下面问题等价：

$$\underset{\substack{h_{ii}\in[0,1]\\ \sum_{i=1}^m h_{ii}=r}}{\text{maximize}} \sum_{i=1}^{m} h_{ii} d_i^2 \tag{7.40}$$

(e) 令 $d_1^2 \geqslant d_2^2 \geqslant \dots d_m^2 \geqslant 0$，求证：式（7.40）的解为 $h_{11} = h_{22} = \dots = h_{rr} = 1$，其他系数为 0。若 $\{d_i^2\}$ 严格有序，求证这个解是唯一的。

(f) 证明：若式（7.38）中最优的 $\boldsymbol{Q}$ 是 $\boldsymbol{U}_1$，则该矩阵为 $\boldsymbol{U}$ 的前 r 列。

习题 7.3 (a) ℓ_1 **范数可当成一个**LP（Linear Programming，线性规划）问题。对于任意的 $\beta \in \mathbb{R}^p$，有

$$\|\beta\|_1 = \max_{\boldsymbol{u}\in\mathbb{R}^p} \sum_{j=1}^{p} u_j\beta_j, \quad \text{其约束为 } \|\boldsymbol{u}\|_\infty \leqslant 1 \tag{7.41}$$

这种关系说明 ℓ_∞ 范数是 ℓ_1 范数的对偶。

(b) ℓ_1 **范数可以当成一个**SDP。对于任意的矩阵 $\boldsymbol{B} \in \mathbb{R}^{m\times n}$，有

$$\|\boldsymbol{B}\|_\star = \max_{\boldsymbol{U}\in\mathbb{R}^{m\times n}} \operatorname{trace}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{B}), \quad \text{其约束为 } \|\boldsymbol{U}\|_{\mathrm{op}} \leqslant 1$$

其中 $\|\boldsymbol{U}\|_{\mathrm{op}}$ 是矩阵 $\boldsymbol{U}$ 的最大奇异值，这被称为**谱范数**或 **ℓ_2算子范数**。这个关系表明谱范数是原子范数的对偶。（提示：使用 $\boldsymbol{Z}$ 的 SVD 和迹算子的循环性质，可将这个关系化简为 (a) 的一个实例）。

(c) 给定矩阵 $\boldsymbol{U} \in \mathbb{R}^{m\times n}$，求证 $\|\boldsymbol{U}\|_{\mathrm{op}} \leqslant 1$ 等价于约束

$$\begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{U} \\ \boldsymbol{U}^{\mathrm{T}} & \boldsymbol{I}_n \end{pmatrix} \succeq 0 \tag{7.42}$$

由于这个约束是一个线性矩阵不等式，可证明最小化原子范数可以看成一个 SDP。（提示：可能会用到 Schur 补。）

习题 7.4 **原子范数的次梯度**。5.2 节定义了次梯度，它是对不可微函数的梯度概念的扩展。

(a) 给定矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，它的秩 $r \leqslant \min(m,n)$，将 $\boldsymbol{A}$ 的奇异值分解记为 $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$。求证，原子范数的次梯度为

$$\partial\|\boldsymbol{A}\|_\star = \left\{\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{W} \,\middle|\, \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{W}\boldsymbol{V} = \boldsymbol{0}, \|\boldsymbol{W}\|_{\mathrm{op}} \leqslant 1\right\} \tag{7.43}$$

(b) 利用 (a) 的结果证明 Soft-Impute 方法（见算法 7.1）的不动点满足式（7.10）的次梯度。

习题 7.5 回顾第 5 章的 Nesterov 广义梯度方法的式（5.21）。求证 Soft-Impute 方法（见算法 7.1）相当于将这个算法应用于式（7.10）。

习题 7.6　构造最大间隔问题（7.21）的解，其形式为 $\hat{\boldsymbol{M}} = \hat{\boldsymbol{A}}_{m\times r}\hat{\boldsymbol{B}}_{r\times n}^{\mathrm{T}}$，其中 $\operatorname{rank}(\boldsymbol{M}) = r < \min(m,n)$。求证这个解不唯一。若假设 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的列数 $r' > r$，则这个扩大问题的解可能不会揭示矩阵 $\boldsymbol{M}$ 的秩。

习题 7.7　对于凸优化问题

$$\underset{\boldsymbol{u}\in\mathbb{R}^p}{\operatorname{maximize}}\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{v}\text{ ，其约束为 }\|\boldsymbol{u}\|_2 \leqslant 1\text{，}\|\boldsymbol{u}\|_1 \leqslant c \tag{7.44}$$

求证一个解为

$$\boldsymbol{u} = \frac{S_\lambda(\boldsymbol{Z}\boldsymbol{v})}{\|S_\lambda(\boldsymbol{Z}\boldsymbol{v})\|_2} \tag{7.45}$$

其中，$\lambda \geqslant 0$ 是使 $\|\mathbf{u}\|_1 \leqslant c$ 最小的正数。

习题 7.8　求证：从无噪声元素中精确填充 $n\times n$ 矩阵 $\boldsymbol{M}$ 时，观测值的个数必须至少为 $N > n\log n$，即便是秩为 1 的矩阵。首先要注意的是，如果无法从 $\boldsymbol{M}$ 的某些行（或列）观察到任意元素，则不可能准确恢复 $\boldsymbol{M}$（即使限制秩为 1 的不相关矩阵）。设采样模式为从矩阵随机均匀地抽取 N 个元素（**有放回抽取**），$\mathcal{F}$ 表示没有元素被观察到的行。

(a) 对于第 j 行（$j = 1,\ldots,p$）令 Z_j 为表示第 j 行是否有元素被观察到的指示变量，定义 $Z = \sum_{j=1}^n Z_j$，求证：

$$\mathbb{P}[\mathcal{F}] = \mathbb{P}[Z > 0] \geqslant \frac{(\mathbb{E}[Z])^2}{\mathbb{E}[Z^2]}$$

（**提示**：这里会用到 Cauchy–Schwarz 不等式。）

(b) 求证：$\mathbb{E}[Z] = n(1-1/n)^N$。

(c) 求证：$\mathbb{E}[Z_iZ_j] \leqslant \mathbb{E}[Z_i]\,\mathbb{E}[Z_j]$，$i \neq j$。

(d) 通过 (b) 和 (c) 证明 $\mathbb{E}\left[Z^2\right] \leqslant n(1-1/n)^N + n^2(1-1/n)^{2N}$。

(e) 使用上面各结果来证明，当 $N > n\log n$ 时，$\mathbb{P}[\mathcal{F}]$ 会远离 0。

习题 7.9　对高维数据采用二次多项式回归是很危险的，这是因为参数数量与维度的平方成比。证明如何将这个问题表示为矩阵回归（见 7.5 节），并提出方案，控制参数爆炸。

习题 7.10　习题 2.11 中给出了针对每个特征的稀疏扰动的回归模型，这相当于采用 Huber 的 ρ 函数健壮回归。健壮 PCA 会有一个类似的结果。

稀疏加低秩的 PCA 为

$$\underset{\boldsymbol{L},\boldsymbol{S}}{\operatorname{minimize}}\ \frac{1}{2}\|\boldsymbol{Z}-(\boldsymbol{L}+\boldsymbol{S})\|_F^2 + \lambda_1\|\boldsymbol{L}\|_\star + \lambda_2\|\boldsymbol{S}\|_1 \tag{7.46}$$

现在考虑 PCA 的健壮版本

$$\underset{\boldsymbol{L}}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{p} \rho(z_{ij} - \ell_{ij}; \lambda_2) + \lambda_1 \|\boldsymbol{L}\|_\star \tag{7.47}$$

其中，

$$\rho(t;\lambda) = \begin{cases} \lambda|t| - \lambda^2/2, & |t| > \lambda \\ t^2/2, & |t| \leqslant \lambda \end{cases} \tag{7.48}$$

是 Huber 损失函数，求证：对于 $\boldsymbol{L}$，问题（7.47）与问题（7.46）有相同的解。

第 8 章　稀疏多元方法

8.1　引　　言

本章将介绍一些流行的多元分析方法，探讨如何将它们稀疏化，即减少特征以便得到解释性更好的解。许多标准多元方法都与某种数据矩阵的奇异值分解有关。因此，对于同样的数据矩阵，不同的稀疏分解会得到相对应的稀疏多元分析。7.6 节的惩罚矩阵分解就是非常好的稀疏分解，因为它给出了左奇异向量和右奇异向量的稀疏版本。

对于一个 $N \times p$ 矩阵 $\boldsymbol{X}$，假设每列的样本均值为 0。$\boldsymbol{X}$ 的主成分可由奇异值分解 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$ 得到，即 $\boldsymbol{V}$ 的列为主成分的方向向量，$\boldsymbol{U}$ 的列为归一化主成分。因此，为了得到稀疏主成分，可以对 $\boldsymbol{X}$ 采用惩罚矩阵分解并让右奇异向量具有稀疏性。与之类似，许多多元方法都可以由适当的惩罚矩阵分解得到。这些方法汇总在表 8-1 中。

表 8-1　在不同的输入矩阵上采用 7.6 节的惩罚矩阵分解得到各种经典多元方法的稀疏版本

输入矩阵	结果
数据矩阵	稀疏 SVD 和主成分
方差–协方差矩阵	稀疏主成分
交叉相乘矩阵	稀疏的典型变量
相异矩阵	稀疏聚类
类意协方差矩阵	稀疏线性判别

8.2　稀疏组成分分析

下面介绍稀疏主成分分析，它是 PCA 的自然延伸，非常适合高维数据。此处先介绍主成分分析。

8.2.1　背景

给定数据矩阵 $\boldsymbol{X}$（大小为 $N \times p$），它由 $\mathbb{R}^p$ 中的 N 个样本 $\{x_1, \ldots, x_N\}$ 构成。主成分分析提供了一组线性近似，其大小由秩 $r \leqslant \min\{p, N\}$ 决定。

有两种不同但等效的方式可以进行主成分分析。第一种方式基于**最大方差**。任

意范数为 1 的向量 $\alpha \in \mathbb{R}^p$ 会得到一维数据投影，即 N 维向量 $\boldsymbol{X}\alpha$①。假设 $\boldsymbol{X}$ 的列已经中心化，则投影数据向量的样本方差为 $\widehat{\mathrm{Var}}(\boldsymbol{X}\alpha) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i^{\mathrm{T}}\alpha)^2$。主成分分析是为了找到具有最大样本方差

$$v_1 = \arg\max_{\|\alpha\|_2=1}\left\{\widehat{\mathrm{Var}(\boldsymbol{X}\alpha)}\right\} = \arg\max_{\|\alpha\|_2=1}\left\{\alpha^{\mathrm{T}}\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{N}\alpha\right\} \tag{8.1}$$

的方向。因此，第一主成分的方向对应样本协方差 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/N$ 的最大特向量，它与总体水平的最大方差有关。详细的介绍参见习题 8.1。图 8-1 给出了这种优化问题的几何示意图。投影 $\boldsymbol{z}_1 = \boldsymbol{X}v_1$ 称为数据 $\boldsymbol{X}$ 的第一主成分，v_1 称为**主成分载荷**，其中 v_1 就是 $\boldsymbol{X}$ 的最大奇异值 d_1 对应的右奇异向量。与之类似，还有 $\boldsymbol{z}_1 = \boldsymbol{u}_1 d_1$，其中 $\boldsymbol{u}_1$ 是最大奇异值 d_1 对应的左奇异向量。

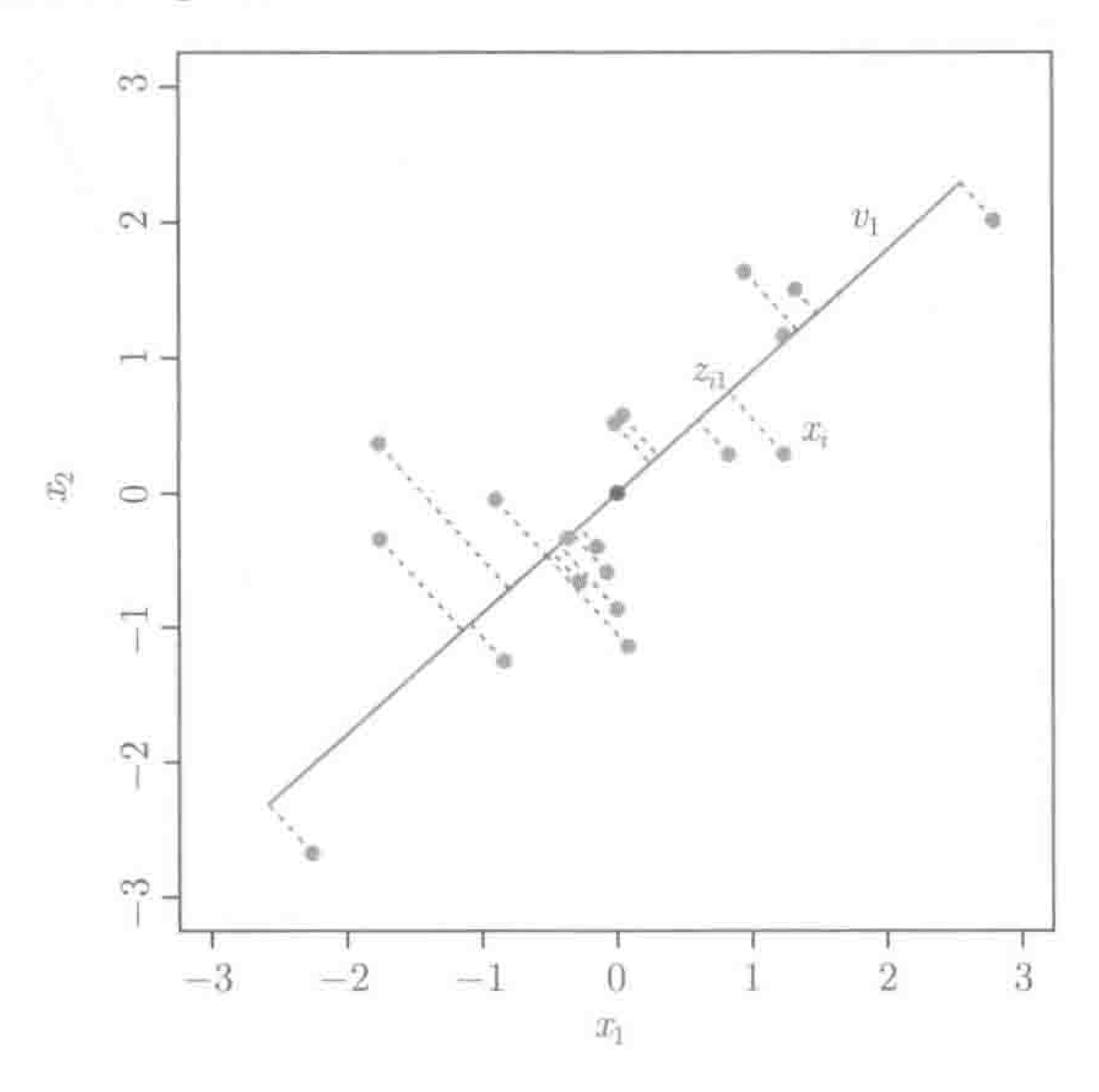


图 8-1　主成分分析的二维示意图，图中给出了数据点集 $x_i = (x_{i1}, x_{x2})$，表示为实心小圆点，第一主成分 $v_1 \in \mathbb{R}^2$。设 $\bar{x} = (\bar{x}_1, \bar{x}_2)^{\mathrm{T}}$ 表示样本均值，这些数据点在直线 $\bar{x} + \lambda^{\mathrm{T}}v_1$ 上的投影方差最大，并且每个点到直线上距离加在一起最小。这里 $z_{i1} = u_{i1}d_1$ 是标量值，表示 x_i 在第一主成分 $\boldsymbol{z}_1$ 中的值

其他主成分方向（特征向量）为 $v_2, v_3, \ldots, v_p$，都有 $\widehat{\mathrm{Var}}(\boldsymbol{X}v_j)$，其约束为 $\|v_j\|_2 = 1$，且 $v_1, \cdots, v_{j-1}$ 与 v_j 正交。由这个性质可得出 $\boldsymbol{z}_{\mathrm{j}}$ 互不相关（见习题 8.2）。事实上，这个过程的第 r 步之后，会得到一个秩为 r 的矩阵，这个矩阵也可以通过求解优

① 本章会用多元方法来处理数据矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，因此约定用粗体表示所有矩阵和 N 维向量，使用普通方式表示 p 维向量。

化问题

$$\boldsymbol{V}_r = \underset{\boldsymbol{A}:\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I}_r}{\arg\max}\ \operatorname{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}) \tag{8.2}$$

得到。习题 7.2 给出了这个性质更详细的介绍。即使这些向量是依次得到，载荷向量的集合 $\boldsymbol{V}_r$ 也能够最大化所有这些集合中的整体方差。

推导主成分分析的第二种方法基于最小化**重构误差**，这种误差与该数据的特定生成模型相关。假设对数据矩阵的行可建立模型 $x_i \approx f(\lambda_i)$，其中 $f(\cdot)$ 定义为

$$f(\lambda) = \mu + \boldsymbol{A}_r\lambda \tag{8.3}$$

该函数会参数化一个 r 维的仿射集，其中 $\mu \in \mathbb{R}^p$ 是位置向量。$\boldsymbol{A}_r \in \mathbb{R}^{p\times r}$ 各列彼此正交，$\lambda \in \mathbb{R}^r$ 是参数向量，与样本有关系。因此，需要选择一组参数 $\{\mu, \boldsymbol{A}_r, \{\lambda_i\}_{i=1}^N\}$，使重构误差

$$\frac{1}{N}\sum_{i=1}^{N}\|x_i - \mu - \boldsymbol{A}_r\lambda_i\|_2^2 \tag{8.4}$$

最小。图 8-1 给出了 PCA 的这种解释。数据若像习题 8.3 那样被预处理过（即 $\mu = 0$），则式（8.4）可以简化为

$$\frac{1}{N}\sum_{i=1}^{N}\left\|x_i - \boldsymbol{A}_r\boldsymbol{A}_r^{\mathrm{T}}x_i\right\|_2^2 \tag{8.5}$$

$\boldsymbol{A}_r$ 会让式（8.5）中的重构误差最小，它可以通过数据矩阵的奇异值分解获得，即计算 $\boldsymbol{X}$ 的 SVD：$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$，则 $\hat{\boldsymbol{A}}_r = \boldsymbol{V}_r$，其中 $\boldsymbol{V}_r$ 由 $\boldsymbol{V}$ 的前 r 个奇异向量列构成。由 $\boldsymbol{Z}_r = \boldsymbol{U}_r\boldsymbol{D}_r$ 的行可以估计 λ_i。因此，最大化仿射表面整体方差相当于最小化样本点到仿射表面的总距离。这是一个**嵌套**求解的过程。这个性质很特殊，不是本章讨论的主要内容。

8.2.2 稀疏主成分

前面已经多次提到载荷向量 $\{v_j\}_{j=1}^r$，这是因为它对解释主成分非常重要。本节将讨论如何得到具有稀疏载荷的主要成分。当特征数量 p 比样本数大时，这种稀疏主成分特别有用。一般来说，当特征数量很大时，最好通过载荷向量来选择较小的相关变量子集。从理论上讲，若 $p \gg N$，则普通 PCA 的效果很差，因为样本协方差的特征向量不会靠近总体特征向量（Johnstone 2001）。对于“P 大，N 小”的应用，主成分的稀疏性会使问题变得适定（well-posed）。这一节会讨论一些获得稀疏主成分的方法，这些方法均基于 lasso 型（ℓ_1）惩罚。与普通 PCA 一样，这里需要 $N \times p$ 的数据矩阵 $\boldsymbol{X}$ 是列中心化的。所提出的方法也会通过最大方差或最小重构误差来建立目标函数。为了便于说明，在此对每一种方法首先讨论秩为 1 的情形，然后在 8.2.3 节讨论更复杂的情形。

1. 基于最大方差的稀疏性

我们先从方差最大化的角度来讨论 PCA 的稀疏性。最自然的修改是对目标函数增加一个 ℓ_0 约束，从而得到目标函数。

$$\underset{\|v\|_2=1}{\text{maximize}}\left\{v^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v\right\}, \quad 其约束为\ \|v\|_0 \leqslant t \tag{8.6}$$

其中 $\|v\|_0=\sum_{j=1}^{P} I[v_j \neq 0]$ 表示向量 v 中非零元素的个数。但这是一个双重非凸问题，因为它要最大化（而非最小化）一个带有组合约束的凸函数。Jolliffe，Trendafilov and Uddin (2003) 提出的 SCoTLASS 方法, 用 ℓ_1 范数代替 ℓ_0 范数，从而得到目标函数。

$$\underset{\|v\|_2=1}{\text{maximize}}\left\{v^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v\right\}, \quad 其中\ \|v\|_1 \leqslant t \tag{8.7}$$

ℓ_1 约束会迫使 v 中的某些元素为 0，从而使其具有稀疏性。虽然 ℓ_1 范数是凸的，但整个问题仍为非凸，而且不适合采用简单迭代算法来进行求解。

有很多方法可以解决这个问题，其中一种为主成分分析的 SVD 版本。重写这个问题，保留非凸状态，可用一种有效的算法来得到局部最优解。对于基于惩罚矩阵的问题（7.28），若 $c_1=\infty$，则关于 $\boldsymbol{u}$ 的约束会不起作用，这会得到优化问题

$$\underset{\|\boldsymbol{u}\|_2=\|v\|_2=1}{\text{maximize}}\left\{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}v\right\}, \quad 其约束为\ \|v\|_1 \leqslant t \tag{8.8}$$

这个问题的优化解 $\hat{v}$ 也是式（8.7）的最优解。式（8.8）是关于 $(\boldsymbol{u}, v)$ 的双凸问题，因此可用交替最小化方法来求解。实际上，这与第 7 章的惩罚矩阵分解（见算法 7.2）类似。算法 8.1 用于求解问题（8.8），它包括如下步骤:

算法 8.1 秩 1 稀疏 PCA 的交替迭代算法

---

1. 设 $v \in \mathbb{R}^p$，有 $\|v\|_2=1$。
2. 一直迭代直到 $\boldsymbol{u}$ 和 v 都足够小。
   (a) 通过 $\boldsymbol{u} \leftarrow \frac{Xv}{\|Xv\|_2}$ 更新 $\boldsymbol{u} \in \mathbb{R}^N$。
   (b) 通过
$$v \leftarrow v(\boldsymbol{\lambda}, \boldsymbol{u})=\frac{\mathcal{S}_\lambda(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})}{\|\mathcal{S}_\lambda(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})\|_2} \tag{8.9}$$
更新 $v \in \mathbb{R}^p$。其中，若 $\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}\|_1 \leqslant t$，则 $\lambda=0$; 否则选择 $\lambda>0$，使得 $\|v(\lambda, \boldsymbol{u})\|_1=t$。

这里的 $\mathcal{S}_\lambda=\operatorname{sgn}(\mathrm{x})(|\mathrm{x}|-\lambda)_+$ 是 λ 处的软阈值算子。习题 8.6 会证明这个算法得到的任意不动点都是式（8.7）的局部最优解。而且，该迭代可被解释为**优化–最大化**，或只是目标函数（8.7）的优化算法。

另一种方法由 d'Aspremont, El Ghaoui, Jordan and Lanckriet（2007）提出，该方法进一步将 SCoTLASS 目标函数松弛成凸规划问题。实际上，这种方法会在半正定矩阵空间上将所求解问题转换成线性优化问题。这样的优化问题也是半定规

划。为了理解该方法，首先重写非凸目标函数（8.7）。利用矩阵迹的性质，可将二次形式 $v^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v$ 写成

$$v^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v = \mathrm{trace}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}vv^{\mathrm{T}}) \tag{8.10}$$

令秩 1 矩阵 $\boldsymbol{M} = vv^{\mathrm{T}}$。约束 $\|v\|_2^2 = 1$ 与线性线束 $\mathrm{trace}(\boldsymbol{M}) = 1$ 等价。约束 $\|v\|_1 \leqslant 1$ 可写为 $\mathrm{trace}(|\boldsymbol{M}|\boldsymbol{E}) \leqslant t^2$，其中 $\boldsymbol{E} \in \mathbb{R}^{p\times p}$ 为全 1 矩阵，$|\boldsymbol{M}|$ 表示对矩阵 $\boldsymbol{M}$ 中各元素取绝对值，因此，可将非凸的 SCoTLASS 目标函数重写为

$$\begin{aligned}&\underset{\boldsymbol{M}\succeq 0}{\mathrm{maximize}}\, \mathrm{trace}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{M})\\ &\text{其约束为}\mathrm{trace}(\boldsymbol{M}) = 1，\mathrm{trace}(|\boldsymbol{M}|\boldsymbol{E}) \leqslant t^2，\text{且 } \mathrm{rank}(\boldsymbol{M}) = 1\end{aligned} \tag{8.11}$$

这样改写后，式（8.11）的最优解会是一个秩为 1 半正定矩阵，如 $\boldsymbol{M} = vv^{\mathrm{T}}$ 。因此，向量 v 是原问题（8.7）的一个优化解。但由于有约束 $\mathrm{rank}(\boldsymbol{M}) = 1$，式（8.11）仍是一个非凸问题，若去掉该约束，可得到一个半定规划问题 (见 d'Aspremont et al. 2007)，即

$$\begin{aligned}&\underset{\boldsymbol{M}\succeq 0}{\mathrm{maximize}}\ \ \mathrm{trace}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{M})\\ &\text{其约束为 } \mathrm{trace}(\boldsymbol{M}) = 1，\mathrm{trace}(|\boldsymbol{M}|\,\boldsymbol{E}) \leqslant t^2\end{aligned} \tag{8.12}$$

这是一个凸规划问题，因此它没有局部最优解，可通过多种标准方法（包括内点法，见 Boyd and Vandenberghe 2004）来得到全局最优解。d'Aspremont et al.（2007）针对特殊情形给出了一种更有效的方法来求解它。

一般情况下，求解式（8.12）这样的 SDP 问题相比求解式（8.8）这种双凸问题，局部最优的计算量要更大。但式（8.12）是一个凸规划问题，有更吸引人的理论保证：如果求解 SDP，并获得一个秩 1 解，则非凸问题 SCoTLASS 的全局最优解实际已经获得。对各类协方差模型可证明，像式（8.12）这样的 SDP 问题，只要样本大小 N 相对于稀疏性和维数充分大（但仍允许 $N \ll p$ ），则解为秩 1 矩阵的概率很高。8.2.6 节会进一步讨论这个问题。对于所有这类情况，SCoTLASS 都可保证找到全局最优解。

2. 基于重构的方法

现在来讨论基于重构误差的稀疏 PCA。在单个稀疏主成分的情况下，Zou, Hastie and Tibshirani（2006）提出了优化问题

$$\underset{\substack{\theta, v\in\mathbb{R}^p\\ \|\theta\|_2=1}}{\mathrm{minimize}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \left\| x_i - \theta v^{\mathrm{T}} x_i \right\|_2^2 + \lambda_1 \|v\|_1 + \lambda_2 \|v\|_2^2 \right\} \tag{8.13}$$

其中，λ_1 和 λ_2 为非负的正则化参数。下面对式（8.13）进行一些说明。

- 如果 $\lambda_1 = \lambda_2 = 0$，则式 (8.13) 的最优解为 $\hat{\theta} = \hat{v} = v_1$，它是 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的最大特征向量，也就是普通 PCA 的解。
- 当 $p \gg N$ 时，如果 $\lambda_2 > 0$，其解唯一。如果设置 $\lambda_1 = 0$，则对于任意 $\lambda_2 > 0$，优化解 $\hat{v}$ 的方向与最大主成分方向一样。
- 在一般情况下，λ_1 和 λ_2 严格为正，ℓ_1 惩罚项的权重 λ_1 会让载荷向量具有稀疏性。

与式 (8.8) 一样，若将 θ 和 v 结合起来看，式 (8.13) 并不是凸问题，但它是双凸问题。固定 θ，将 v 当成变量来最小化目标函数，这便等价于弹性网问题（见习题 4.2），可以有效求解。另外，固定 v，将 θ 当成变量来最小化目标函数，则可直接得到解

$$\boldsymbol{\theta} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}}{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}\|_2} \tag{8.14}$$

其中 $\boldsymbol{z}_i = v^{\mathrm{T}}\boldsymbol{x}_i, i = 1, \ldots, N$（见习题 8.8）。总的来说，此算法相当有效，但没有算法 8.1 那样简单，该算法只涉及软阈值。

事实证明，原始 SCoTLASS 式 (8.7) 和基于回归的目标函数 (8.13) 密切相关。注意，可将基于秩 1 的式 (8.13) 看成带约束的优化问题的拉格朗日形式，即

$$\underset{\|v\|_2=\|\theta\|_2=1}{\text{minimize}} \left\|\boldsymbol{X} - \boldsymbol{X}v\theta^{\mathrm{T}}\right\|_F^2, \quad \text{其约束为 } \|v\|_1 \leqslant t \tag{8.15}$$

如果增加一个额外的 ℓ_1 约束 $\|\theta\|_1 \leqslant t$，则可证明，这个优化问题的结果等于 SCoTLASS 式 (8.7)（即习题 8.7）的求解。这里可以运用算法 8.1。注意，增加 ℓ_1 约束相当自然，因为它只是在式 (8.15) 中让约束对称化。

8.2.3　秩大于 1 的解

8.2.1 节对标准主成分分析提出了一种顺序求解方法，这种方法会连续求解秩 1 问题，并让每个解与其他解正交。这种顺序方法也可用来求解秩大于 1 的问题 (8.2)。

如何得到稀疏解呢？SCoTLASS 算法也采用这种顺序方法，其中秩为 k 的候选解都要与前面秩小于 k 的解正交。但是这里的顺序方法通常不能用来求解秩大于 1 的问题。

对于式 (8.8) 的稀疏 PCA，可以采用第 7 章多因子惩罚的矩阵分解 (7.3) 进行求解。给定秩 1 解 $(\boldsymbol{u}_1, v_1, d_1)$，可简单计算残差 $\boldsymbol{X}' = \boldsymbol{X} - d_1\boldsymbol{u}_1v_1$，然后对 $\boldsymbol{X}'$ 采用求解式 (8.8) 的秩 1 算法[①]。这样做不会让主成分 $\{(\boldsymbol{u}_1, d_1), (\boldsymbol{u}_2d_2), \ldots, (\boldsymbol{u}_kd_k)\}$ 正交，也不会让载荷向量 $\{v_1, v_2, \ldots, v_k\}$ 稀疏。但在实践应用中，有些主成分会正交。

① 没有稀疏约束，此过程将得到普通主成分序列。

这里有一个微妙的问题：在稀疏 PCA 的情形下，向量 $\{v_1, v_2, \dots, v_k\}$ 是否还有正交性，这一点并不清楚，因为正交性可能与稀疏不一致。另外还需说明，强制正交性可能会导致解的稀疏变少。类似的问题在稀疏编码中也会出现，8.2.5 节会就此讨论。

有趣的是，式（8.8）可以增加约束，让向量 $\boldsymbol{u}_j$ 彼此正交，而向量 v_j 没有这样的约束。这样的修改可提高解的可解释性，同时仍可让 v_j 稀疏。具体而言，可以得到优化问题

$$\begin{aligned}&\underset{\boldsymbol{u}_k, v_k}{\text{maximize}}\left\{\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{X} v_k\right\}，\text{其约束为 } \|v_k\|_2 \leqslant 1，\|v_k\|_1 \leqslant c \text{ 且}\\ &\|\boldsymbol{u}_k\|_2 \leqslant 1，\quad \boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{u}_j = 0，\quad j = 1, \dots, k-1\end{aligned} \tag{8.16}$$

在 v_k 给定的情况下，$\boldsymbol{u}_k$ 的解为

$$\boldsymbol{u}_k = \frac{\boldsymbol{P}_{k-1}^{\perp} \boldsymbol{X} v_k}{\left\|\boldsymbol{P}_{k-1}^{\perp} \boldsymbol{X} v_k\right\|_2} \tag{8.17}$$

其中，$\boldsymbol{P}_{k-1}^{\perp} = \boldsymbol{I} - \sum_{i=1}^{k-1} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}$ 投影在 $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_{k-1}$ 所张成的正交补空间上。将式（8.17）代替秩 1 投影 $\boldsymbol{u} \leftarrow \frac{\boldsymbol{X} v}{\|\boldsymbol{X} v\|_2}$，可以得到算法 8.1 的多因子版本。

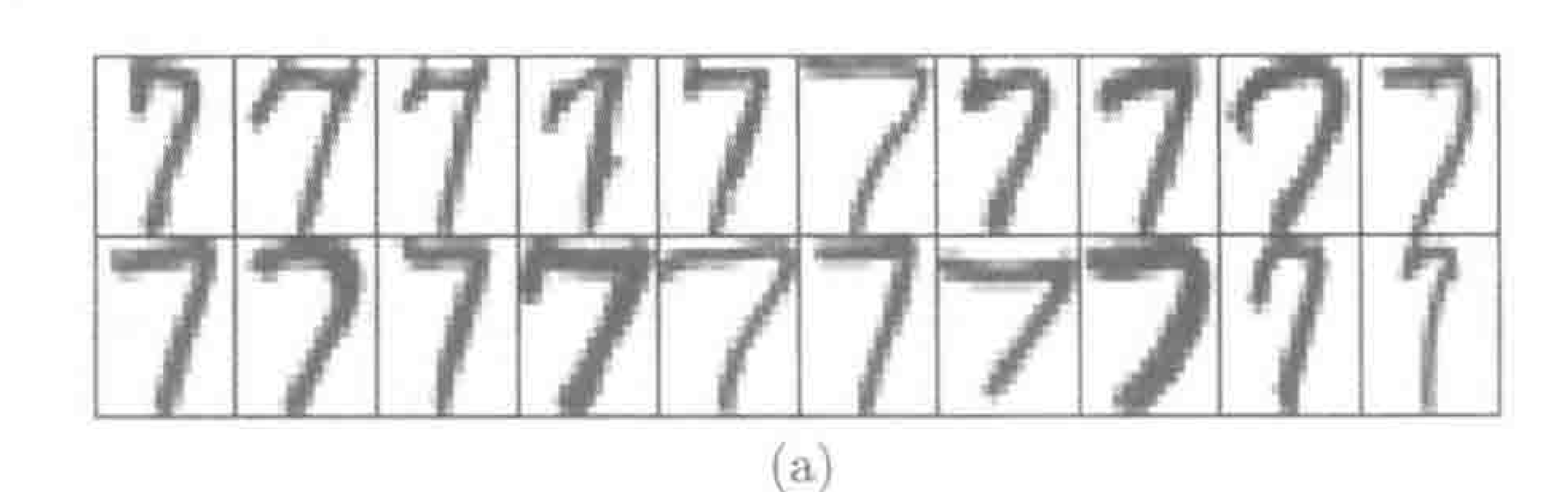
Zou et al.（2006）提出的方法（8.13）可以看成是最小化目标函数

$$\frac{1}{N} \sum_{i=1}^{N} \left\|x_i - \boldsymbol{\Theta} \boldsymbol{V}^{\mathrm{T}} x_i\right\|_2^2 + \sum_{k=1}^{r} \lambda_{1k} \|v_k\|_1 + \lambda_2 \sum_{k=1}^{r} \|v_k\|_2^2 \tag{8.18}$$

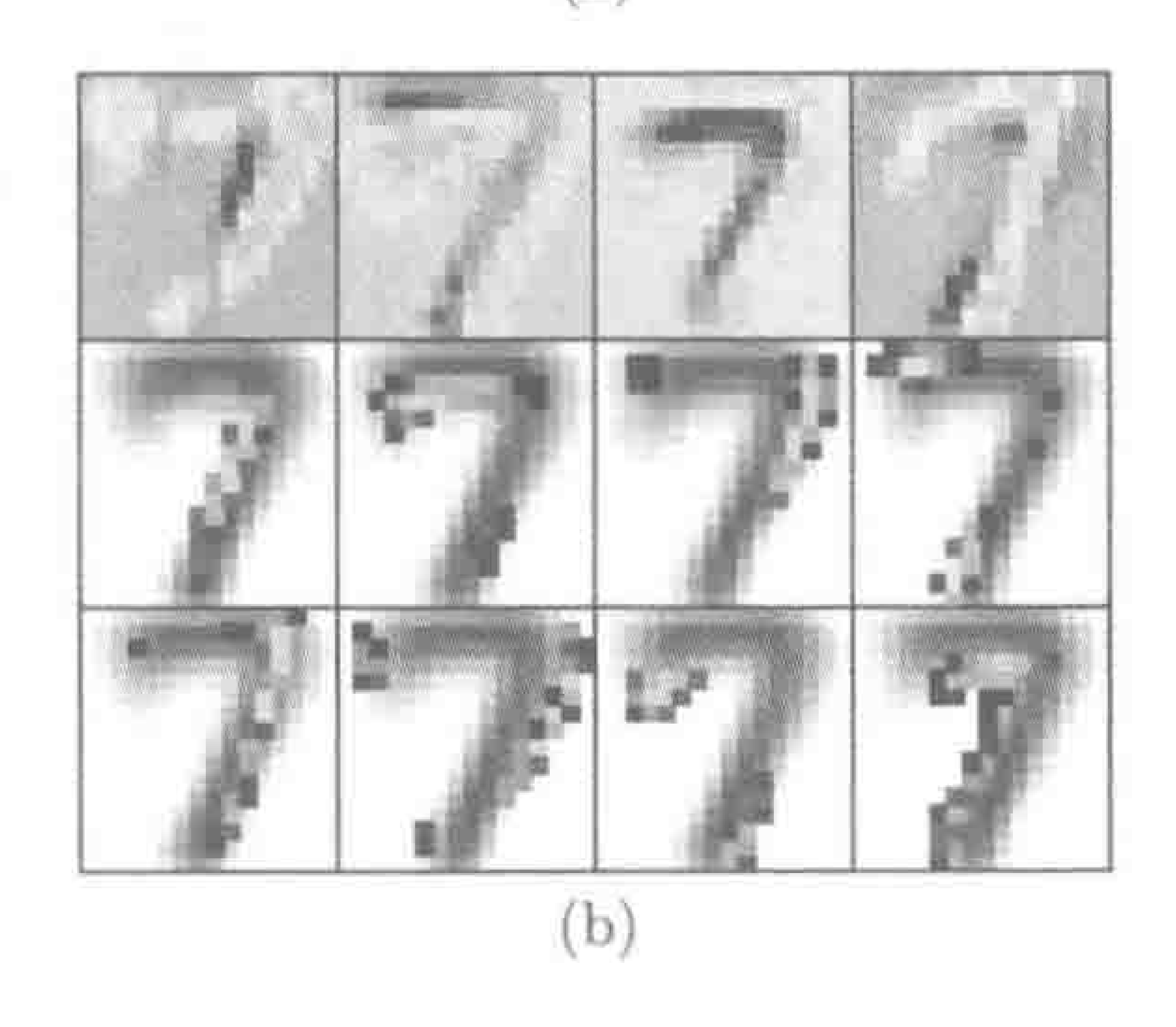
的多秩情形，其中 $\boldsymbol{\Theta}^{\mathrm{T}} \boldsymbol{\Theta} = \boldsymbol{I}_{r \times r}$。$\boldsymbol{V}$ 是一个 $p \times r$ 的矩阵，它的列为 $\{v_1, \dots, v_r\}$，$\boldsymbol{\Theta}$ 也是一个 $p \times r$ 的矩阵。虽然将 $\boldsymbol{\Theta}$ 和 $\boldsymbol{V}$ 结合起来看，式（8.18）并不是凸问题，但它是一个双凸问题。固定 $\boldsymbol{\Theta}$，基于 $\boldsymbol{V}$ 最小化目标函数，等价于单独解 r 个弹性网问题，求解会非常有效。另外，固定 $\boldsymbol{V}$，基于 $\boldsymbol{\Theta}$ 最小化目标函数，等价于 Procrusters 问题，用 SVD 就可以求解（见习题 8.10）。这样不断进行交替迭代，直到收敛到局部最优解。

稀疏 PCA 的示例应用

下面在手写数字数据集上展示稀疏主成分。这个训练集的样本数 $N = 664$，每个样本都是灰度图像。每幅图像的大小为 16×16 像素，因此，数据矩阵 $\boldsymbol{X}$ 的大小为 664×256。图 8-2a 为训练样本的例子，而 b 是稀疏主成分的结果，并将其与标准 PCA 进行对比。b 最上一行显示前 4 个标准主成分，它们能解释大约 50%的变化。为了提高可解释性，需限制稀疏主成分的载荷为非负。要做到这一点，只需用非负的软阈值算子 $\mathcal{S}_\lambda^{+}(x) = (x - \lambda)_+$ 替换算法 8.1 中的软阈值算子 $\mathcal{S}_\lambda(x)$。



(a)



(b)

图 8-2　(a) 取自邮政编码数据库中的手写数字 7。(b) 上面一行，手写数字 7 的前 4 个主成分（不同颜色表示不同的载荷：负载荷为黄色；正载荷为蓝色）；下面两行，前 8 个稀疏主成分，载荷被限制为正数。这些稀疏主成分叠加到普通的数字 7 上，以提高解释性（见彩插）

中间和最后一行显示了前 8 个稀疏主成分，它们也可解释 50% 的变化。当用更多个成分来解释相同的变化量时，各个成分就都变得简单，而且可能更具有可解释。例如，第 2 个和第 6 个稀疏成分似乎捕获了书写这些数字者所产生的 “缺口” 风格，见图 8-2 a 中左上角图像。

8.2.4　基于 Fantope 投影的稀疏 PCA

Vu, Cho, Lei and Rohe（2013）提出了另一种与稀疏 PCA 相关的方法。令 $\boldsymbol{S}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/N$，这个方法可以求解半定规划

$$\underset{\boldsymbol{Z}\in\mathcal{F}^p}{\text{maximize}}\left\{\operatorname{trace}(\boldsymbol{S}\boldsymbol{Z})-\lambda\|\boldsymbol{Z}\|_1\right\} \tag{8.19}$$

其中凸集 $\mathcal{F}^p=\{\boldsymbol{Z}:\mathbf{0}\preceq\boldsymbol{Z}\preceq\boldsymbol{I},\operatorname{trace}(\boldsymbol{Z})=p\}$ 称为 Fantope。当 $p=1$ 时，在 $\mathcal{F}^p$ 上的谱范数冗余，式（8.19）会退化为 d'Aspremont et al.（2007）提出的方法。当 $p>1$ 时，虽然式（8.19）中的惩罚仅表明解的逐元素稀疏，但是可以证明（Lei and Vu 2015），在合适的条件下，这个解能一致地选择主导特征向量的非零项。

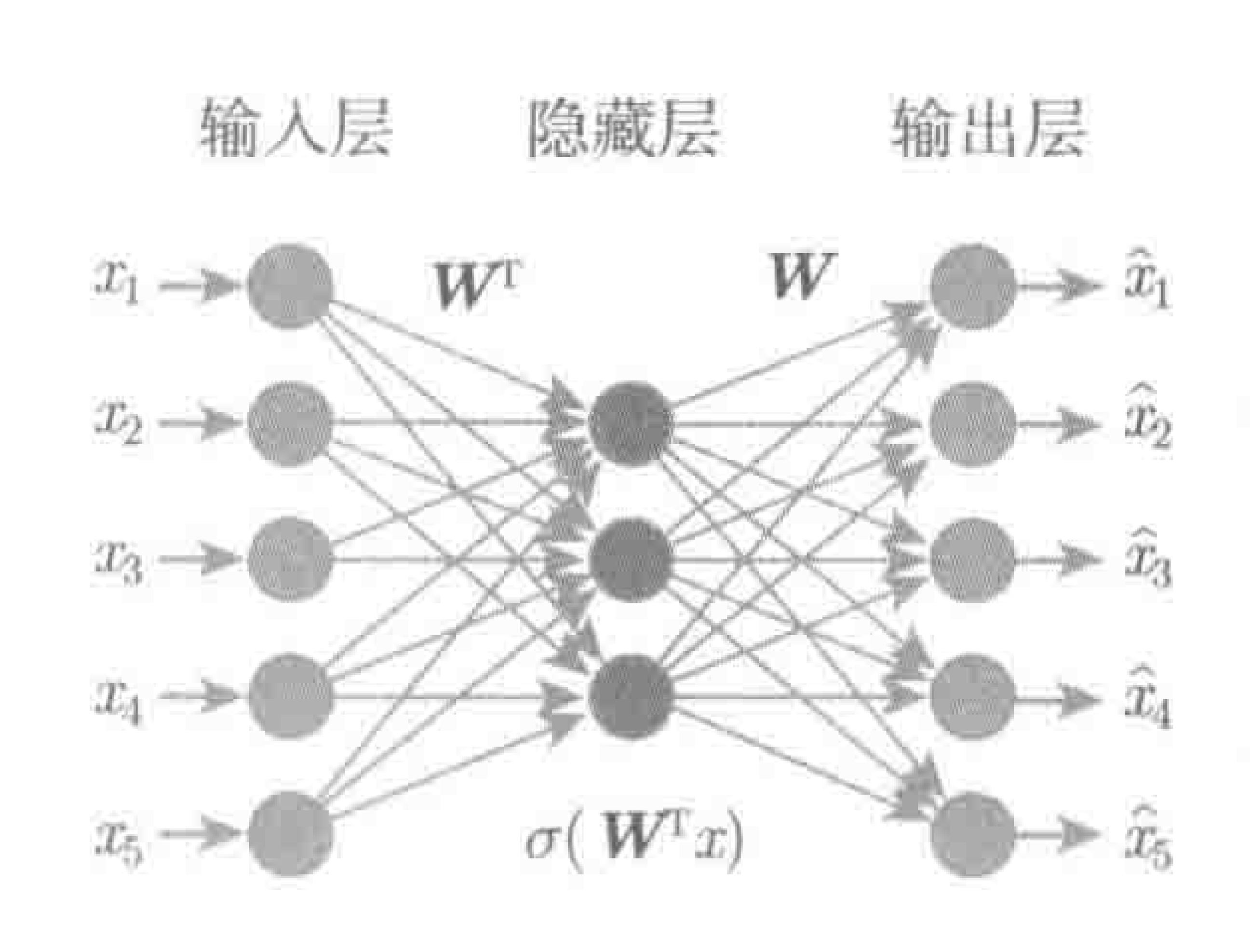
8.2.5　稀疏自编码和深度学习

神经网络的参考文献认为，自编码（autoencoder）源于主成分的概念。图 8-3 为这一观点提供了简单的示意图，这种想法基于重建误差，很像式（8.13）。自编码

基于 $p\times m$（$m<p$）的权重矩阵 $\boldsymbol{W}$，会为输入向量 x 创建 m 个线性组合。每个线性组合会传递给一个非线性函数 σ，通常 σ 会取 $\sigma(t)=1/(1+\mathrm{e}^{-t})$，则输出层为 $\boldsymbol{W}h(x)=\boldsymbol{W}\sigma(\boldsymbol{W}^{\mathrm{T}}x)$[①]。给定输入向量 $x_i(i=1,\ldots,N)$，通过求解（非凸）优化问题

$$\underset{\boldsymbol{W}\in\mathbb{R}^{m\times p}}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{N}\|x_i-\boldsymbol{W}h(x_i)\|^2\right\} \tag{8.20}$$

来得到权重矩阵 $\boldsymbol{W}$。如果 σ 取恒等函数，则 $h(\boldsymbol{x})=\boldsymbol{W}^{\mathrm{T}}x$，式（8.20）的解与主成分分析的等价，即 $\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}=\boldsymbol{V}_m\boldsymbol{V}_m^{\mathrm{T}}$，其中 $\boldsymbol{V}_m$ 是 $p\times m$ 矩阵，由前 m 个主成分载荷构成（见习题 8.12）。这种神经网络有一个瓶颈，中间层会约束 $\boldsymbol{W}$ 的秩，并让 $\boldsymbol{W}$ 去学习结构。



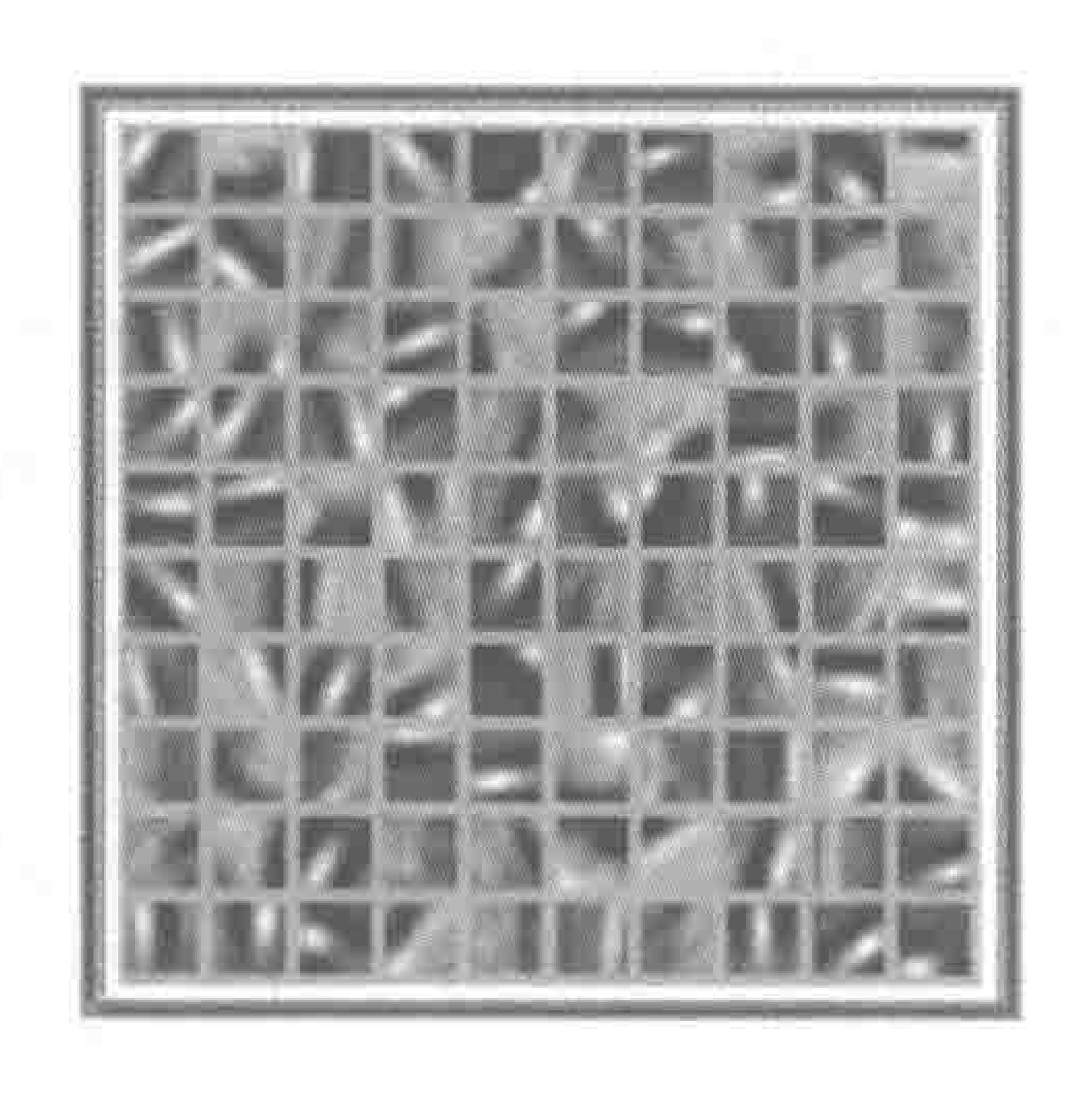


图 8-3 左图：自编码的网络表示，它用于非线性主成分的无监督学习。中间层的隐藏单元创建了一个瓶颈，并学习输入的非线性表示。输出层的权重矩阵是输入层的权重矩阵的转置，因此该网络试图用这种限制性表示再现输入。右图：在一幅图像的建模任务中，用图像来表示估计到的 $\boldsymbol{W}$ 的列

对高维信号（比如图像）建模时，向量 x_i 可能表示一幅（子）图像的像素。$\boldsymbol{W}$ 的列表示学习到的图像形状字典，$h(x_i)$ 试着用这组基来表示 x_i。在这种情形下，中间层看起来是好像无用的约束，因为各种不同形状很可能出现在同一幅图像中。可以让系数 $h(x_i)$ 稀疏以替代这种约束，这就是所谓的**稀疏编码**（sparse coding，Olshausen and Field 1996）。为了便于理解，首先考虑线性情形。$m>p$，优化问题为

$$\underset{\boldsymbol{W}\in\mathbb{R}^{p\times m},\{s_i\}_1^N\in\mathbb{R}^m}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{N}\left\{\|x_i-\boldsymbol{W}s_i\|_2^2+\lambda\|s_i\|_1\right\}\right\}\text{，其约束为 }\|\boldsymbol{W}\|_F^2\leqslant 1 \tag{8.21}$$

其中单个 $\boldsymbol{s}_i$ 通过 ℓ_1 惩罚项来得到稀疏性。并不需要限制 $\boldsymbol{W}$ 的列不相关，只是让它的 Frobenius 范数小于等于 1。习题 8.13 更详细地介绍了如何估计稀疏线性编码

① 在实践应用中，线性组合还会包括偏置项，这里为简单起见，在此省略这部分内容。

式 (8.21)。目标函数可用交替迭代算法来求解。图像所得到的 $\boldsymbol{W}$ 通常会像图 8-3 右图一样，每个 x_i 都是一幅图像的向量化版本。每一个子图像可用 $\boldsymbol{W}$ 的一列（编码本）来表示。一个图像通过 $\boldsymbol{W}$ 元素的稀疏叠加来建模。有如下几种方式可将该公式推广成用于深度学习的现代稀疏编码方法（Le et al. 2012）:

- 使用多个隐藏层会得到层次字典结构;
- 使用比 σ 函数计算速度更快的非线性函数，例如 $\sigma(t) = t_+$;
- 直接对式 (8.20) 中系数 $h(x_i)$ 进行稀疏化。

这些模型可以由（随机）梯度下降来拟和，通常非常大的数据库（比如图像数据库）一样可以使用基于大规模处理器集群的分布式计算。

稀疏自编码的一个重要用途是**预训练** (pretraining)。当拟和有标签数据的监督神经网络时，通常要首先利用无标签数据得到自编码器，然后将得到的权重作为初始值来拟和监督神经网络（Erhan et al. 2010）。由于神经网络的目标函数为非凸函数，因而这些起始权重可显著改善最终解的质量。另外，如果还有无标签数据，可在预训练阶段采用这些数据得到自编码。

8.2.6 稀疏 PCA 的一些理论

下面简要介绍一下标准主成分分析为什么会在高维情形 ($p \gg N$) 下失败，以及为什么一些结构性的假设（如主成分稀疏）必不可少。（稀疏）PCA 行为可以用**尖协方差模型** (spiked covariance model) 来研究，这种模型的 p 维协方差矩阵形如

$$\boldsymbol{\Sigma} = \sum_{j=1}^{M} \omega_j \theta_j \theta_j^{\mathrm{T}} + \sigma^2 \boldsymbol{I}_{p\times p} \tag{8.22}$$

其中，向量 $\{\theta_j\}_{j=1}^M$ 彼此正交，正权重 $\omega_1 \geqslant \omega_2 \geqslant \ldots \geqslant \omega_M > 0$。向量 $\{\theta_j\}_{j=1}^M$ 是协方差矩阵的前 M 个特征向量，对应的特征值为 $\left\{\sigma^2 + \omega_j\right\}_{j=1}^M$。

设有 N 个独立同分布的样本 $\{x_i\}_{i=1}^N$，它们服从协方差为 $\boldsymbol{\Sigma}$ 的零均值分布。标准 PCA 会得到样本协方差矩阵 $\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^N x_i x_i^{\mathrm{T}}$ 的前 M 个特征向量的 $\{\theta_j\}_{j=1}^M$。在典型情形下，维度 p 保持固定，而样本大小 $N \to +\infty$，则样本协方差收敛于总体协方差，即主成分分析是一致估计。更相关的高维数据分析是 p 和 N 都趋于无穷，即 $p/N \to c \in (0, \infty)$，并且 M 和特征值都不变[①]。在这种情形下，样本特征向量（或主成分）并不收敛到总体特征向量 $\{\theta_j^{(p)}\}_{j=1}^M$。事实上，如果信噪比 ω_j/σ^2 充分小，样本特征向量会渐近正交于总体特征向量。这种情况是由尖协方差模型 (8.22) 中 $p - M$ 维的噪声引起的，这会在 $N \ll p$ 时淹没信号。Johnstone and Lu (2009) 对这种现象进行了准确的解释。

① 需要明确：对每个 $j = 1, \ldots, M$，有总体特征向量序列 $\left\{\theta_j^{(p)}\right\}$，而且保持固定信噪比 ω_j/σ^2 与 (p, N) 独立。

高维 PCA 在特征向量没有任何结构的情况下效果很差，因此需要作出额外的假设。许多研究者已经证明稀疏性如何在 $N \ll p$ 时得到一致性估计的主成分。Johnstone and Lu（2009）提出了一个两阶段算法，它会阈值化样本协方差矩阵的对角线元素，以便分离出最大方差坐标，然后在降维后的空间进行 PCA。他们证明，即使 p/N 远大于零，该方法也具有一致性，但仅允许 p 以样本大小的多项式函数进行增涨。Amini and Wainwright（2009）分析了对角元素阈值化和 SCoTLASS 问题（8.7）的半定规划松弛（8.12）的变量选择性质。对于尖协方差模型（8.22），其单个主导特征向量为 k 稀疏，可表明当且仅当 $N \asymp k^2 \log p$ 时，对角元素阈值化法（Johnstone and Lu 2009）可成功恢复稀疏主导特征向量。在这种样本数量下，SDP 松弛也能正确执行变量选择，在一定情形下，用较少样本也会成功。Amini and Wainwright （2009）已证明，在样本数比 $N \asymp k^2 \log p$ 更少的情形下，没有一种方法能成功，穷举所有子集也不行。

其他研究人员采用 ℓ_2 或相关规范来研究特征空间（eigenspace）的估计。Paul and Johnstone（2008）提出了增强 SPCA 算法，Johnstone and Lu（2009）提出了细化的两阶段法。Birnbaum, Johnstone, Nadler and Paul（2013）分析了这个算法并得出：采用 ℓ_q 球，该算法可得到弱稀疏向量模型的极小极大率（minimax rate）。Vu and Lei（2012）证明了稀疏 PCA 问题的极小极大下界，并且证明这些下界可通过计算样本协方差（有 ℓ_q 球约束）的最大特征值来得到。Ma（2010，2013）和 Yuan and Zhang（2013）研究了稀疏 PCA 与 power 方法（用于计算特征向量的迭代算法）相结合的算法。当 $M = 1$ 时，Ma（2013）的方法与算法 8.1 基本相同，唯一的区别是在软阈值步骤中使用了一个固定的 λ，而不是对后者采用可变选择来求解问题的限制版本。

8.3 稀疏典型相关分析

典型相关分析（Canonical Correlation Analysis，CCA）将主成分分析的思想扩展到了两个数据矩阵。假设有数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，它们的大小分别 $N \times p$ 和 $N \times q$，而且它们的列都进行了中心化。给定两个向量 $\beta \in \mathbb{R}^p$，$\theta \in \mathbb{R}^q$，它们定义了两个数据集的一维投影，即变量（N 维向量）$\boldsymbol{X}\beta$ 和 $\boldsymbol{Y}\theta$。典型相关分析就是选择 β 和 θ 来使两组变量之间有最大相关性。

具体而言，$\boldsymbol{X}\beta$ 和 $\boldsymbol{Y}\theta$ 之间的样本协方差为

$$\widehat{\operatorname{Cov}}(\boldsymbol{X}\beta, \boldsymbol{Y}\theta) = \frac{1}{N}\sum_{i=1}^{N}(x_i^{\mathrm{T}}\beta)(y_i^{\mathrm{T}}\theta) = \frac{1}{N}\beta^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\theta \tag{8.23}$$

其中，x_i 和 y_i 分别表示矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的第 i 行。CCA 会求解目标函数

$$\underset{\beta\in\mathbb{R}^p,\theta\in\mathbb{R}^q}{\text{maximize}}\left\{\widehat{\text{Cov}}(\boldsymbol{X}\beta,\boldsymbol{Y}\theta)\right\}，\ \text{其约束为}\ \widehat{\text{Var}}(\boldsymbol{X}\beta)=1\ ,\widehat{\text{Var}}(\boldsymbol{Y}\theta)=1 \tag{8.24}$$

解集 (β_1,θ_1) 称为第一典型向量，相应的线性组 $\boldsymbol{z}_1=\boldsymbol{X}\beta_1$和$\boldsymbol{s}_1=\boldsymbol{Y}\theta_1$ 称为第一个典型变量。随后的变量对可通过约束找到，使得所得变量与前面的变量不相关。所有解都能由矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 的广义 SVD 得到（见习题 8.14）。

若样本大小 N 严格小于 (p,q) 的最大值，则典型相关分析失败。这种情况会得到一个退化问题，找到相关性为 1 的无意义解。一种避免样本协方差矩阵 $\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 和 $\frac{1}{N}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 奇异化的方法是增加额外的限制。比如，岭正则化方法（ridge regularization）对每个样本协方差矩阵加上 λ 乘以一个单位矩阵，其中 λ 大于零，习题 8.17 会进一步讨论这个问题。另一种方法是只取样本协方差矩阵的对角线元素，下面采用这种方法。

对式（8.24）中的 β和θ 采用 ℓ_1 约束，可得到稀疏典型向量，这会得到目标函数

$$\underset{\beta,\theta}{\text{maximize}}\left\{\widehat{\text{Cov}}(\boldsymbol{X}\beta,\boldsymbol{Y}\theta)\right\}$$
$$\text{其约束为}\ \text{Var}(\boldsymbol{X}\beta)=1,\|\beta\|_1\leqslant c_1,\text{Var}(\boldsymbol{Y}\theta)=1,\|\theta\|_1\leqslant c_2 \tag{8.25}$$

注意，由 ℓ_1 约束得到稀疏典型向量的形式有两种：(1) 让 ℓ_1 约束小于某个正数，见式（8.25）；(2) 添加相应的拉格朗日项。这个问题有诸多解答，标准的 CCA 问题（8.24）可通过交替最小二乘回归来求解（见习题 8.14 至 8.17）。同样，式（8.25）可通过交替弹性网求解，习题 8.19 会详细讨论。

当 $N>\max(p,q)$ 时，式（8.25）很有用，但在高维情形中可能会失败，就像前面介绍的一样。此问题可用脊状（ridging）的单个协方差矩阵来解决，同时还可以采用交替弹性网回归求解。当维度非常高（如基因组问题的情形）时，主要考虑 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的交叉相乘，而内部协方差（即 $\boldsymbol{X}$ 各列之间的协方差，$\boldsymbol{Y}$ 各列之间的协方差）是多余的参数，它们是用来估计方差的。在这种情况下，可对特征进行归一化，然后令内部协方差矩阵为单位矩阵。这样得到问题：

$$\underset{\beta,\theta}{\text{maximize}}\left\{\widehat{\text{Cov}}(\boldsymbol{X}\beta,\boldsymbol{Y}\theta)\right\}$$
$$\text{其约束为}\ \|\beta\|^2\leqslant 1,\|\theta\|^2\leqslant 1,\|\beta\|_1\leqslant c_1,\|\theta\|_1\leqslant c_2 \tag{8.26}$$

该目标函数与第 7 章讨论的惩罚矩阵分解（7.6）有相同的形式，只是这里将数据矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 作为输入。因此，可直接利用算法 7.2，采取交替软阈值来计算式（8.26）的解。

高阶的稀疏典型变量能够从高阶的 PMD 成分得到，如算法 7.3 那样，在求解完成后得到残差，然后将这个过程应用于剩下的部分。

示例：Netflix 电影评分数据

下面通过在 Netflix 电影评分数据中使用稀疏 CCA 来解释稀疏 CCA 的性质。如 7.3.1 节所述，整个数据集有 17 770 部电影，480 189 位用户。用户对（约 1%）

电影的评分为 1~5。这个例子的 p 和 N 都取 500，即选择评分最多的 500 部电影和客户，并用影片的均值来估计缺失值。

在这 500 部电影中，确定了动作片（59 部）和爱情片（73 部），其余影片丢弃。对这些数据使用稀疏 CCA，了解每个客户对爱情片评分后，也对动作片评分的可能性。这里将 500 位用户随机分成大小相等的训练集和测试集，并在训练集上应用稀疏 CCA。为了得到可解释性，这里让权重向量为非负。表 8-2 是第一稀疏成分中权重为正的电影。也许影迷能说出这些电影的评分为什么具有关联性。例如，这些动作片与《终结者》相比可能相对平淡一些。图 8-4 给出了测试集中 7 部动作片的平均得分，然后给出每位用户对 16 部爱情片的平均得分。相关性系数为 0.7，这相当地高。因此，对于一个给定的用户，可以从他对 16 部爱情片的平均评分来预测他对 7 部动作片的平均评分，反之亦然。

表 8-2 小 Netflix 数据集：第一稀疏成分中权重为正所对应的动作片和爱情片

	动作片	
《生死时速》	《反恐特警组》	《黑衣人 II》
《速度与激情》	《深入敌后》	《霹雳娇娃》
《空中监狱》		
	爱情片	
《偷听女人心》	《人鬼情未了》	《居家男人》
《保镖》	《特工佳丽》	《风月俏佳人》
《修女也疯狂》	《辣身舞》	《落跑新娘》
《新婚告急》	《曼哈顿灰姑娘》	《贴身情人》
《律政俏佳人 2》	《女孩梦三十》	《新岳父大人》
《律政俏佳人》		

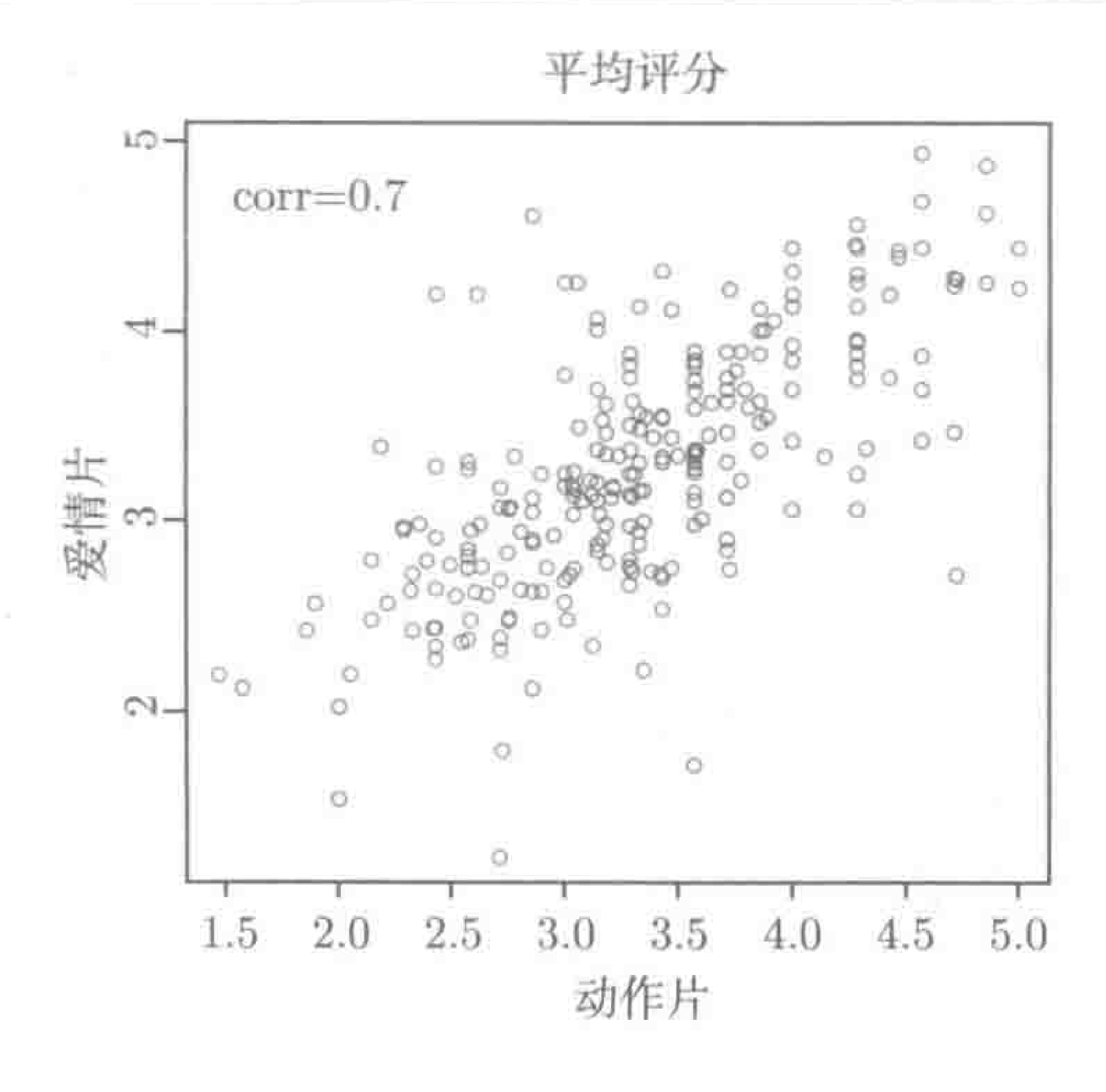


图 8-4 将稀疏典型相关分析应用于 Netflix 电影评分数据的子集上。图中所示为测试数据上 7 部动作片的平均得分。第一稀疏 CCA 成分的非零权重所对应的 16 部爱情片取平均评分即可得到这个结果

8.4 稀疏线性判别分析

线性判别分析（Linear Discriminant Analysis，LDA）是一种重要的分类方法。有很多方法可以实现稀疏线性判别分析。从某种意义上讲，至少有三种不同方式可以改进经典判别分析，它们分别是：标准理论模型、Fisher 的类间–类内方差准则和最优评分。此外，在高维（$p \gg N$）情形下，估计类内协方差需要某种形式的正则化，不同的估计形式会得到不同的稀疏 LDA 方法。

8.4.1 标准理论和贝叶斯规则

设输出变量 G 取 $\{1,2,\ldots,K\}$ 中的某个值，特征 $\boldsymbol{X} \in \mathbb{R}^p$，$f_k(x)$ 是 $G=k$ 时 $\boldsymbol{X}$ 的类条件密度函数，π_k 是第 k 类的先验概率，并有 $\sum_{k=1}^K \pi_k = 1$。应用贝叶斯法则，有

$$\Pr(G=k|\boldsymbol{X}=\boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{\ell=1}^K \pi_\ell f_\ell(\boldsymbol{x})} \tag{8.27}$$

进一步假设每类密度函数为多元高斯函数 $N(\mu_k, \boldsymbol{\Sigma}_w)$，具体形式为

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_w\right|^{1/2}} \mathrm{e}^{-\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_w^{-1}(x-\mu_k)} \tag{8.28}$$

$\boldsymbol{\Sigma}_w$ 为协方差矩阵。对于第 k 类和 ℓ 类，它们的后验概率的对数比率为

$$\begin{aligned}\log\frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} &= \log\frac{f_k(x)}{f_\ell(x)} + \log\frac{\pi_k}{\pi_\ell} \\ &= \log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^{\mathrm{T}}\boldsymbol{\Sigma}_w^{-1}(\mu_k-\mu_\ell) + x^{\mathrm{T}}\boldsymbol{\Sigma}_w^{-1}(\mu_k-\mu_\ell)\end{aligned} \tag{8.29}$$

这个式子是 x 的线性方程。因此，第 k 类和 ℓ 类之间的决策界（即对于所有 x，都有 $\Pr(G=k|\boldsymbol{X}=x)=\Pr(\mathrm{G}=\ell|\boldsymbol{X}=\mathrm{x})$）为 $\mathbb{R}^p$ 的一个超平面。这种表述对任何类都成立，因此所有的决策界都是线性的。如果将 $\mathbb{R}^p$ 按第 1 类、第 2 类等划分成区域，则这些区域会由超平面分开。

式（8.29）说明这类 LDA 模型也是线性逻辑斯蒂回归模型，二者之间的差别仅在于参数估计的方法不一样。逻辑斯蒂回归使用条件二项式/多项式似然，而 LDA 估计基于 X 和 G 的联合似然（Hastie et al. 2009，Chapter4）。从式（8.29）可以得到线性判别函数（linear discriminant function）

$$\delta_k(x) = x^{\mathrm{T}}\boldsymbol{\Sigma}_w^{-1}\mu_k - \frac{1}{2}\mu_k^{\mathrm{T}}\boldsymbol{\Sigma}_w^{-1}\mu_k + \log\pi_k \tag{8.30}$$

该函数给出了决策规则的等价描述，这会得到分类函数 $G(x)=\operatorname{argmax}_{k\in\{1,\ldots,K\}}\delta_k(x)$。

在实际应用中，高斯类条件分布的参数未知。但给定 N 个带类标签的训练样本 $\{(x_1,g_1),\ldots,(x_N,g_N)\}$，可按如下方法来估计这些参数。设 C_k 表示 $g_i=k$ 的下标索引集，$N_k=|C_k|$ 表示第 k 类样本的数量。可得到 $\hat{\pi}_k=N_k/N$，并且有

$$\hat{\mu}_k=\frac{1}{N_k}\sum_{i\in C_k}x_i \tag{8.31a}$$

$$\hat{\boldsymbol{\Sigma}}_w=\frac{1}{N-K}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^{\mathrm{T}} \tag{8.31b}$$

需注意，$\hat{\boldsymbol{\Sigma}}_w$ 是所有类内协方差相加后的无偏估计。

在高维（$p>N$）情形下，样本类内协方差矩阵 $\hat{\boldsymbol{\Sigma}}_w$ 会成为奇异矩阵，所以必须进行正则化。有很多方法可以实现正则化。本节后面会介绍基于二次正则化的方法（Hastie, Buja and Tibshirani 1995）。

在维度非常高时，可以假设特征不相关，这很实用。这样的假设会使 $\hat{\boldsymbol{\Sigma}}_w$ 成为对角矩阵，从而得到**朴素贝叶斯**分类器，也称为**对角化线性判别分析**（见习题 8.20）。设 $\hat{\sigma}_j^2=s_j^2$ 是第 j 个特征的类内方差，因此被估计的分类规则可简化为

$$\hat{G}(x)=\underset{\ell=1,\ldots,K}{\arg\min}\left\{\sum_{j=1}^{p}\frac{(x_j-\hat{\mu}_{j\ell})^2}{\hat{\sigma}_j^2}-\log\hat{\pi}_k\right\} \tag{8.32}$$

这就是著名的**最近中心规则**（nearest centroid rule）。

8.4.2 最近收缩中心

分类规则（8.32）通常涉及所有特征。当 p 很大时，人们希望通过这些特征的一个子集就能得到所需信息。对模型进行再参数化并进行稀疏惩罚就能得到这样的子集。更具体而言，假设将第 k 类均值向量分解成相加形式 $\mu_k=\bar{x}+\alpha_k$，其中 $\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$ 是整体的均值向量，$\alpha_k\in\mathbb{R}^p(k=1,\ldots K)$ 为第 k 类的反差，并且满足 $\sum_{k=1}^{K}\alpha_k=0$。下面考虑 ℓ_1 正则化的优化问题

$$\underset{\alpha_k\in\mathbb{R}^p,k=1,\ldots,K}{\text{minmize}}\left\{\frac{1}{2N}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j=1}^{p}\frac{(x_{ij}-\bar{x}_j-\alpha_{jk})^2}{s_j^2}+\lambda\sum_{k=1}^{K}\sum_{j=1}^{p}\frac{\sqrt{N_k}}{s_j}|\alpha_{jk}|\right\}$$

$$\text{其约束为}\ \sum_{k=1}^{K}\alpha_{jk}=0\,,\ j=1,\ldots,p \tag{8.33}$$

基于 α_{jk} 的解会简化成与具体类别相关反差的软阈值。具体而言，可定义反差为

$$d_{jk}=\frac{\widetilde{x}_{jk}-\bar{x}_j}{m_k s_j} \tag{8.34}$$

其中 $\widetilde{x}_{jk}=\dfrac{1}{N_k}\sum_{i\in C_k}x_{ij}$，$\widetilde{x}_j$ 表示整体均值向量 $\bar{x}$ 的第 j 个分量，[①] $m_j^2=\frac{1}{N_k}-\frac{1}{N}$，然后采用软阈值算子

$$d'_{jk}=\mathcal{S}_\lambda(d_{jk})=\mathrm{sgn}(d_{jk})(|d_{jk}|-\lambda)_+ \tag{8.35a}$$

可得到收缩中心估计

$$\hat{\mu}'_{jk}=\bar{x}_j+m_k s_j d'_{kj} \tag{8.35b}$$

最后，在最近中心规则（8.32）中使用由收缩中心估计的 μ_{jk}。

假设给定特征 j，对第 k 个类，通过软阈值将 d'_{jk} 设为 0。该特征不参与最近中心规则（8.32）。采用这种方式，最近收缩中心过程会实现自动特征选择。此外，一个特征可能对于某些类有 $d'_{jk}=0$，因此它只会对这些类起作用。

最近收缩中心分类器对高维的分类问题（比如针对基因和蛋白质数据的分类）非常有用。一些可用的软件（Hastie, Tibshirani,Narasimhan and Chu 2003）包括了一些附加特性：为每个 s_j 增加一个小的常量 s_0，当 s_j 接近 0 时，反差可以变得稳定。这里举例说明了各种收敛率。

图 8-5 给出了该算法用于某些淋巴瘤（Lymphoma）癌数据的结果（Hastie，Tibshirani，Narasimhan and Chu 2003）。这些数据由 59 个淋巴瘤病人的 4026 个基因表达式度量组成。样品分为弥漫性大 B 细胞淋巴瘤（Diffuse LargeB-cell Lymphoma，DLBCL）、滤泡性淋巴瘤（Follicular Lymphoma，FL），以及慢性淋巴细胞性淋巴瘤（Chronic Lymphocytic Lymphoma，CLL）。

数据被分成有 39（27,5,7）个样本的训练集和有 20 个样本的测试集。基因通过层次聚类来组织。除了 79 以外，其他所有基因都会收缩到零。请注意，较小的类有较大偏差，最大类 DLBCL 基本上可以决定整个均值。下面会将基于最近收缩中心的分类器与 Fisher 线性判别分析的稀疏版本进行比较。

8.4.3 Fisher 线性判别分析

另一种稀疏判别分析源于 Fisher 判别框架，其主要思想为：数据的低维投影具有分类能力。虽然这些投影主要用于可视化，但是也可以在所生成的子空间上进行高斯分类。

设 $\boldsymbol{X}$ 是 $N\times p$ 观测矩阵 [②]，并假定每列为一个特征，而且这些列的均值已被归一化为零。这样的观测矩阵要找到一个低维投影，使得**类间**方差除以**类内**方差的值尽量大。与前面一样，用 $\hat{\boldsymbol{\Sigma}}_w$ 表示类内协方差矩阵，$\hat{\mu}_k$ 表示第 k 类的样本均值（也称为中心）。类间协方差矩阵 $\hat{\boldsymbol{\Sigma}}_b$ 是这些中心的协方差矩阵，由此可得到

① m_k 是被归一化的常量，它基于分子的方差，可将 d_{jk} 作为一个 t 统计量。

② 观测矩阵，即训练数据构成的矩阵。—— 译者注

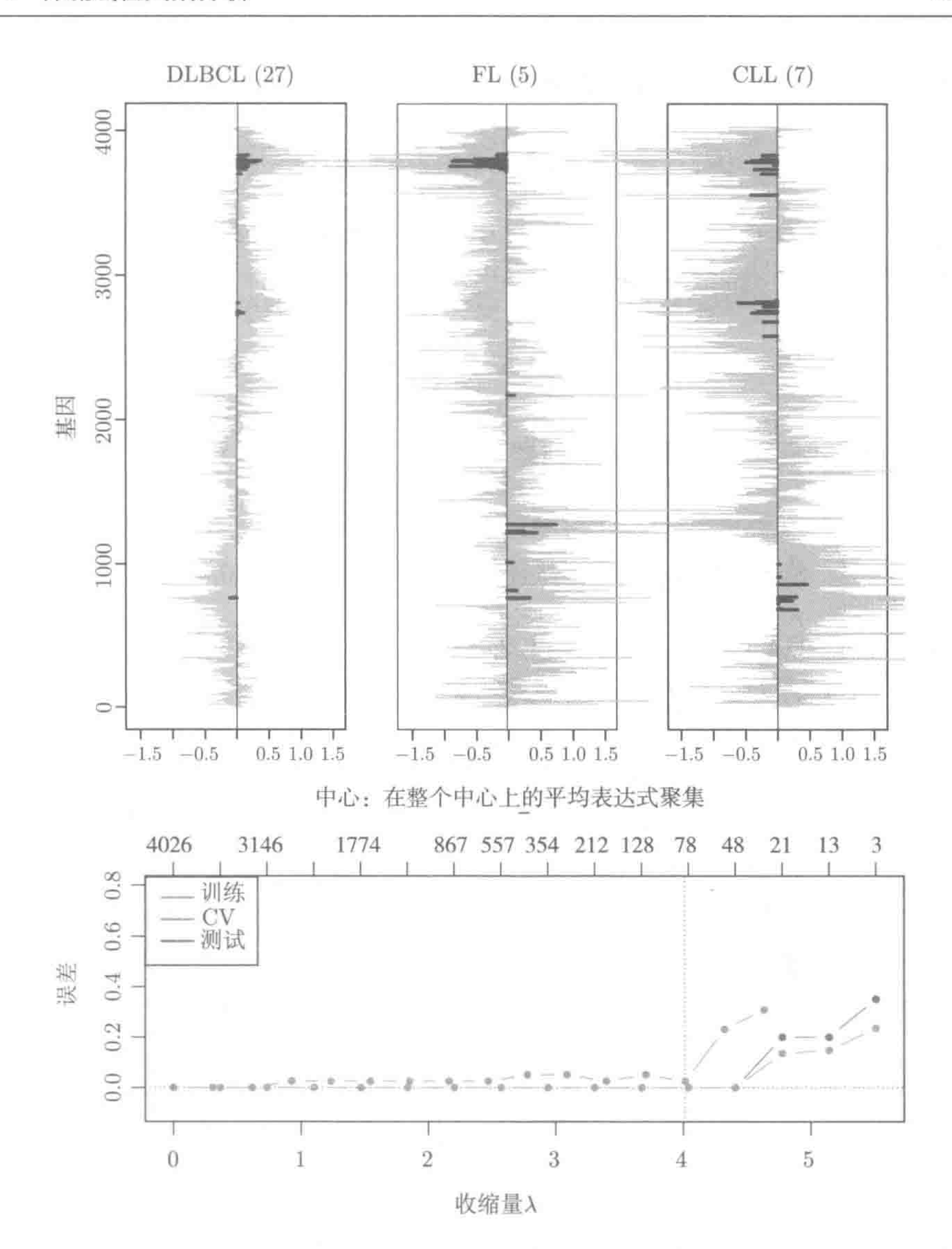


图 8-5 在某些淋巴瘤数据上基于最近收缩中心分类的结果，有三个类。顶部曲线图显示了每个基因的特定类的平均表达式（灰线条），及其收缩版本（黑线条）。大多数基因都缩小到了总体均值（在 0 那里）。下图给出了训练误差、交叉验证误差，以及测试误分类误差与收缩阈值 λ 之间的函数关系。所选择的模型包括 79 个基因，测试误差为 0

$$\hat{\boldsymbol{\Sigma}}_b = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_k \hat{\mu}_k^{\mathrm{T}} \tag{8.36}$$

将这些看成是质量为 π_k 的多元观测，则有

$$\hat{\boldsymbol{\Sigma}}_t = \frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \hat{\boldsymbol{\Sigma}}_b + \hat{\boldsymbol{\Sigma}}_w \tag{8.37}$$

现在假定 $\hat{\boldsymbol{\Sigma}}_w$ 是满秩矩阵（即 $p \leqslant N$），后面会介绍非满秩的情况。对于线性组合 $\boldsymbol{z} = \boldsymbol{X}\beta$，Fisher 的类间方差与类内方差之比为

$$R(\beta) = \frac{\beta^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\beta}{\beta^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_w\beta} \tag{8.38}$$

最大化这个公式即可。可通过求解问题

$$\begin{aligned}&\text{maximize}\left\{\beta^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\beta\right\}\\&\text{使得 } \beta^T\hat{\boldsymbol{\Sigma}}_w\beta \leqslant 1\text{，并且对所有}\ell < k\text{有}\beta^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_w\beta_\ell = 0\end{aligned} \tag{8.39}$$

来求解 Fihser LDA 问题，其中 $k = 1, 2, \ldots, \min(K-1, p)$。虽然问题（8.39）通常会写成不等式约束，但也可以采用等式约束。如果 $\hat{\boldsymbol{\Sigma}}_w$ 是满秩矩阵，则这两种情形等价。解 $\hat{\beta}_k$ 称为第 k 个**判别向量**，$\boldsymbol{z}_k = \boldsymbol{X}\hat{\beta}_k$ 是相应的判别变量。注意 LDA 本质上是在类中心上的主成分，但使用归一化度量与类内方差有关（Hastie et al. 2009，Chapter4）。实际应用并不需要依次求解这个问题，因为这里可像 PCA 那样，用一次特征分解得到所有解。前 k 个判别向量就是矩阵 $\hat{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\Sigma}}_b$ 的前 k 个特征向量。

Witten and Tibshirani（2011）提出了一种稀疏化目标函数（8.39）的方法，即求解目标函数

$$\underset{\beta}{\text{maximize}}\left\{\beta^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\beta - \lambda\sum_{j=1}^{p}\hat{\sigma}_j\,|\beta_j|\right\}\text{，其约束为 }\beta^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_w\beta \leqslant 1 \tag{8.40}$$

其中，$\hat{\sigma}_j^2$ 是 $\hat{\boldsymbol{\Sigma}}_w$ 的第 j 个对角元素，$\tilde{\boldsymbol{\Sigma}}_w$ 是对 $\hat{\boldsymbol{\Sigma}}_w$ 的正定估计。这将产生第一个稀疏判别向量 $\hat{\beta}_1$，其稀疏程度由所选的 λ 决定。接下来先从 $\hat{\boldsymbol{\Sigma}}_b$ 中删除当前解，再依次求解式（8.40），详细介绍参见参考文献。

经过正则化的内类协方差矩阵 $\tilde{\boldsymbol{\Sigma}}_w$ 的参数根据具体情况来选择。某些问题（例如数据为图像时）可以选择 $\tilde{\boldsymbol{\Sigma}}_w$ 来进行空间平滑。可令 $\tilde{\boldsymbol{\Sigma}}_w = \hat{\boldsymbol{\Sigma}}_w + \boldsymbol{\Omega}$，其中，$\boldsymbol{\Omega}$ 会惩罚附近的空间差异值。Hastie, Tibshirani and Buja et al.（1994）和 Hastie et al.（1995）提出的柔性征罚判别分析方法（flexible and penalized discriminant analysis）便基于这种观点。在稀疏情形下，8.4.4 节的最佳得分方法很方便实现。其他情形只需要 $\tilde{\boldsymbol{\Sigma}}_w$ 为 $\hat{\boldsymbol{\Sigma}}_w$ 的正定估计。因此，这里可以得到一个岭化版本：$\tilde{\boldsymbol{\Sigma}}_w = \hat{\boldsymbol{\Sigma}}_w + \varepsilon\operatorname{diag}(\hat{\boldsymbol{\Sigma}}_w)$，$\varepsilon > 0$。

在一种特别有趣的简单情形下，$\tilde{\boldsymbol{\Sigma}}_w$ 为对角矩阵，即 $\mathrm{diag}(\hat{\boldsymbol{\Sigma}}_w)$。这样一来，问题（8.40）可转换为惩罚的矩阵分解，这种分解会用在协方差矩阵 $\hat{\boldsymbol{\Sigma}}_b$ 中，并可用算法 7.2 完成。对于 $K=2$ 的情形，这种方法能得到与最近收缩中心相似的解，详细的介绍可参见 Witten and Tibshirani（2011, Section 7.2）。对于两个以上的类，这两种方法会不一样。最近收缩中心产生每个类和总体平均值之间的稀疏反差，而稀疏 LDA 方法（8.40）会针对更一般的类反差得到稀疏判别向量。区别在下面的例子中介绍。

示例：有 5 个类的模拟数据

这里创建两个情形来对比稀疏判别分析（8.40）和最近收缩中心方法。图 8-6 给出了在两个不同模拟数据集上采用最近收缩中心分类器所得的结果。这两种情形的样本数为 $N=100$，每类有 20 个样本，每个样本的特征 $p=1000$，总共有 $K=5$ 类。

(1) 在第一种情形中，类别 1 的前 10 个特征有两个单位的正反差。类别 3 的特征 11~20 有两个单位正反差，而类别 5 的特征 21~30 为两个单位的负反差。类别为 1、3、5 的前 10 个特征就可以描述这些类。

(2) 在第二种情形中，类别 3~5 的特征 1~10 都有一个单位的正反差（这是相对于类别 1 和 2 而言）；而类别 2 的特征 11~20 有一个单位的正反差，类别 1 的特征 11~20 有一个单位的负反差。因此，相对于类别 1~2 而言，前 10 个特征对类别 3~5 具有判别性，而特征 11~20 对类别 1~2 具有判别性。

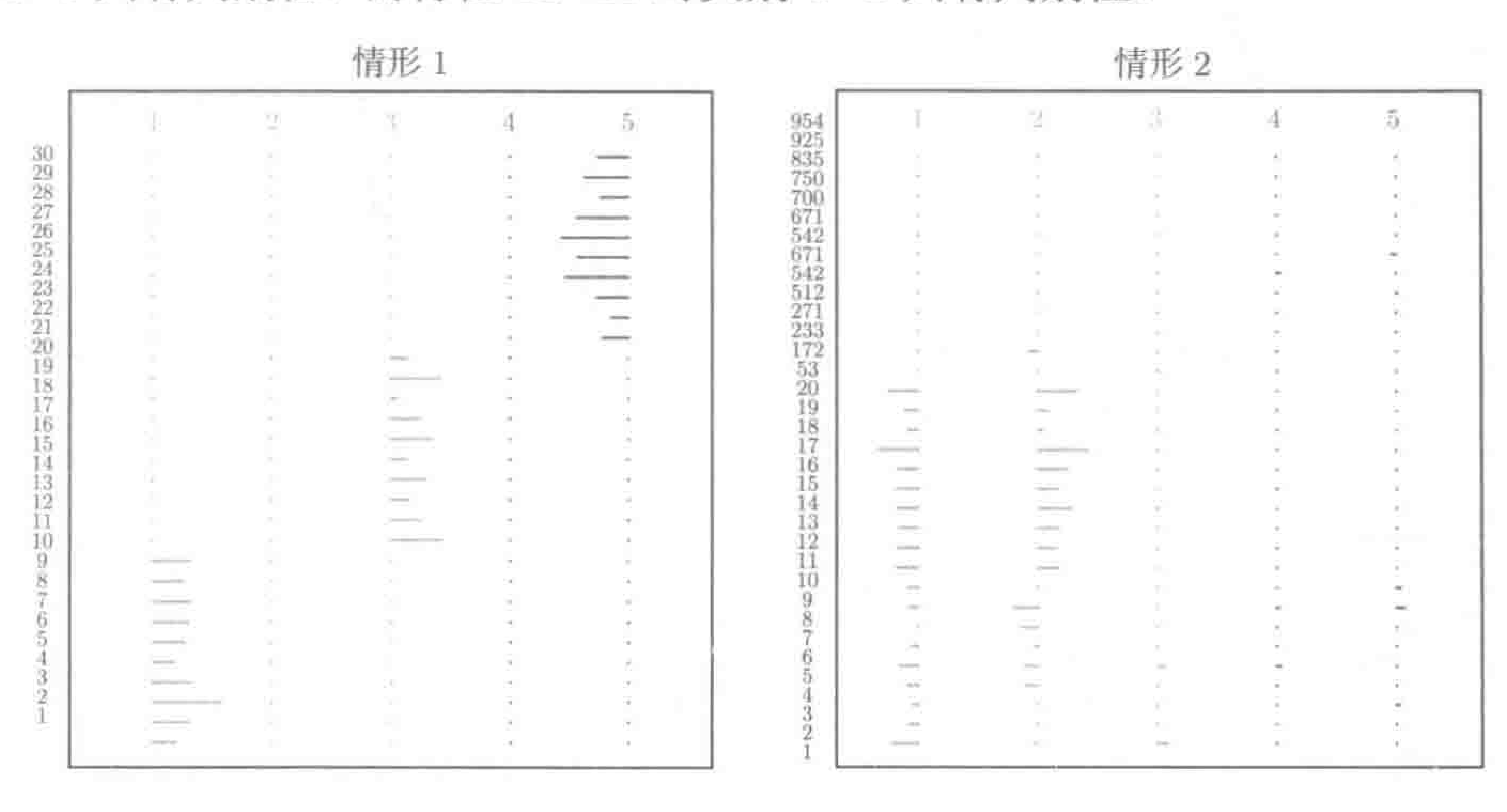


图 8-6 最近收缩中心分类器应用于两个不同模拟数据集上的结果。这里给出了估计的非零特征（每幅图中的行）。每个水平线段的长度对应反差的大小，反差为正则线段朝右，否则线段朝左

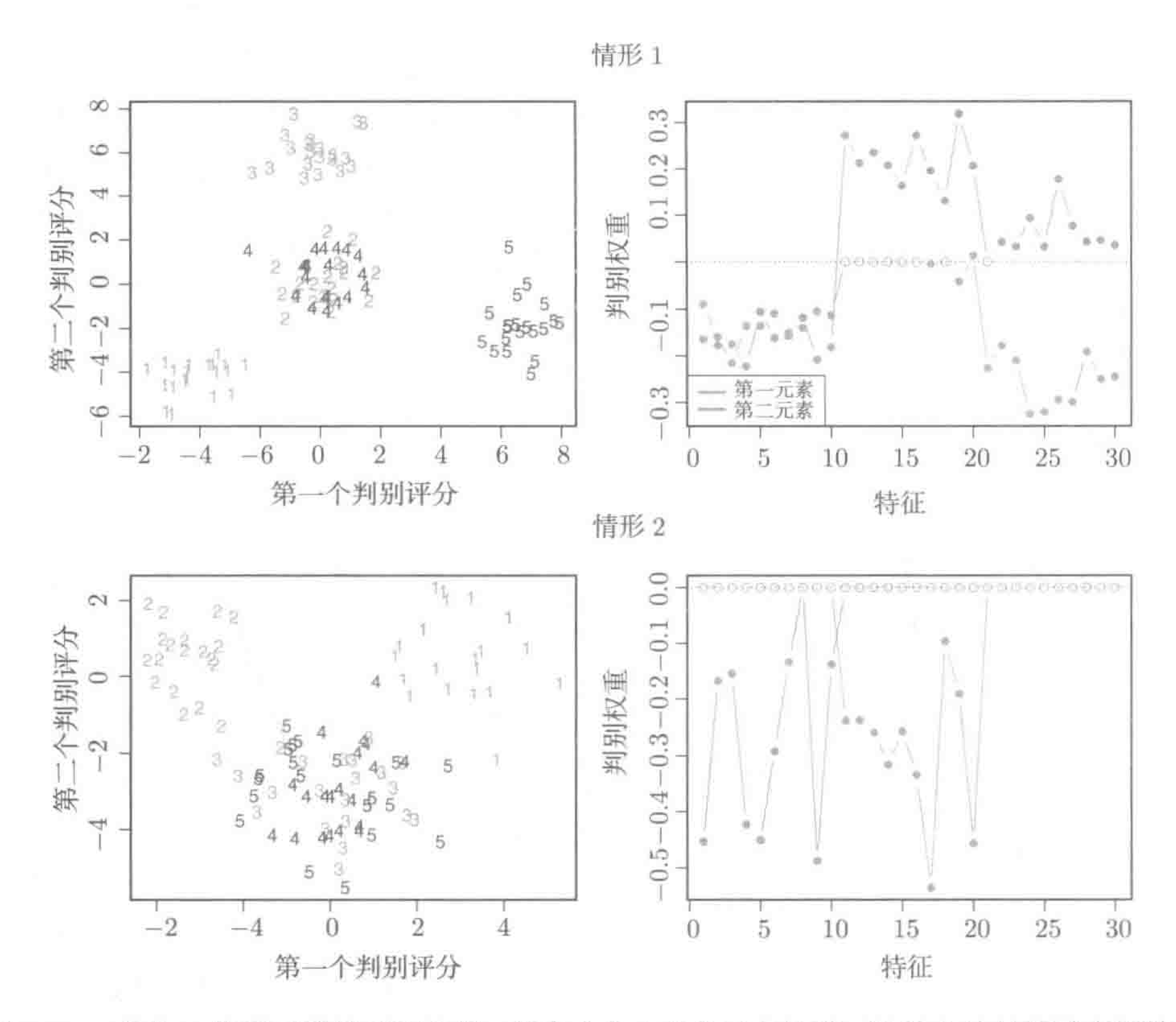


图 8-7　秩为 2 的稀疏线性判别分析，其类内协方差为对角矩阵。这种方法的两种应用情形与图 8-6 一样（顶部为情形 1，底部为情形 2）。左边的两幅图是投影在前两个稀疏判别向量上的分类示意图。右边两幅图为判别权重（也称载荷）

在使用最近收缩中心分类器时，要采用交叉验证来选择收缩参数，并给出具有非零反差的特征。水平线段的长度与反差大小成正比，线段向右表示反差为正，否则反差为负。从左图可知，最近收缩中心能清楚地揭示数据的结构，而右图没有做到这一点。图 8-7 给出了秩为 2 的稀疏线性判别分析的结果，两种情形的类内协方差都为对角矩阵。第一种情形（上面两幅图）下，判别投影可完全将类别 1、3 和 5 与其他类别分开。但三条判别载荷信息被合并成 2 个向量，因此右上角的图看起来很模糊。当然可使用大于 $K = 2$ 的稀疏成分，这在本例中是有益的，但若需要高阶判别，这种方法就有问题。第二种情形非常适合采用稀疏 LDA，在此可以很好地从测试中将类别 1 和 2（左下图）区分开来，并能针对这种分类来提示特征所起的作用（右下图）。

8.4.4　最佳评分

线性判别分析推导的第三种方法称为**最佳评分**。它是对多元线性回归问题的

改进，即通过最优方式选择输出类的编码，下面将详细介绍这种方法。假设用二值指示矩阵 $\boldsymbol{Y}$（其大小为 $N \times K$）对样本进行编码，即

$$y_{ik} = \begin{cases} 1, & \text{第 } i \text{ 个样本属于} k \text{类} \\ 0, & \text{其他} \end{cases}$$

使用这种符号，最佳评分分别针对 $k = 1, \ldots, K$ 来依次求解问题

$$\begin{aligned} (\hat{\beta}_k, \hat{\theta}_k) = & \underset{\beta_k \in \mathbb{R}^p, \theta_k \in \mathbb{R}^K}{\arg\min} \left\{ \frac{1}{N} \|\boldsymbol{Y}\theta_k - \boldsymbol{X}\beta_k\|_2^2 \right\} \\ & \text{使得 } \theta_k^{\mathrm{T}} \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \theta_k = 1 \text{ 且} \theta_k^{\mathrm{T}} \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \theta_j = 0, \ j = 1, 2, \ldots, k-1 \end{aligned} \tag{8.41}$$

可证明这个问题的优化解 $\hat{\beta}_k$ 是 Fisher 线性判别法（8.39）的解（Breiman and Ihaka1984, Hastie et al. 1995）。这种等价性并不令人惊奇。基于 $\boldsymbol{X}$ 的线性回归，其输出 $\tilde{y} = \boldsymbol{Y}\theta$ 为二值结果（θ 为任意编码），它所得到的系数与 2 分类的线性判别分析一样（最多相差一个比例因子），详细信息参见 Hastie et al.（2009）中的习题 4.2。若超过两个类，$\boldsymbol{X}$ 上的回归 $\tilde{\boldsymbol{y}}_\ell$ 会因每个类分配到的数值分数 $\theta_{\ell k}$ 而不同。当线性回归与通过评分选择优化而得到的线性判别分析等价时，可以从线性回归得到解，同问题（8.41）一样。

与稀疏判别分析方法一样，这里可以为式（8.41）增加一个 ℓ_1 惩罚项，得到优化问题

$$\begin{aligned} & \underset{\beta_k \in \mathbb{R}^p, \theta_k \in \mathbb{R}^K}{\text{minimize}} \left\{ \frac{1}{N} \|\boldsymbol{Y}\theta_k - \boldsymbol{X}\beta_k\|_2^2 + \beta_k^{\mathrm{T}} \boldsymbol{\Omega} \beta_k + \lambda \|\beta_k\|_1 \right\} \\ & \text{使得 } \theta_k^{\mathrm{T}} \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \theta_k = 1 \text{ 且} \theta_k^{\mathrm{T}} \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \theta_j = 0, \ j = 1, 2, \ldots, k-1 \end{aligned} \tag{8.42}$$

（见 Leng 2008 和 Clemmensen, Hastie, Witten and Ersboll 2011。）除了增加非负正则参数 λ 的 ℓ_1 惩罚项外，还需要增加一个二次惩罚项，由半正定矩阵 $\boldsymbol{\Omega}$ 定义。若 $\boldsymbol{\Omega} = \gamma \boldsymbol{I}$，则与弹性网等价。如果 ℓ_1 惩罚项对应的参数 λ 充分大，则会得到稀疏的判别向量。若 $\lambda = 0$，则最小化式（8.42）相当于 Hastie et al.（1995）提出的惩罚判别分析（penalized discriminant analysis）。虽然该问题为非凸（因为二次约束），但是可以通过交替最小化来得到局部最优。β 可利用弹性网求解。事实上，对于最佳评分问题（8.41），如果有任何凸惩罚用于判别向量，则很容易用交替最小化来求解该问题。此外，这种方法与稀疏 Fisher LDA（8.40）之间有密切联系。实际上，如果取 $\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_w + \boldsymbol{\Omega}$，则它们实质上是等价的。由于非凸性，这里用了"实质"。只能说，一个问题的稳定点（stationary point）也是另一个问题的稳定点（详见习题 8.22）。

根据问题性质的不同，可选择用 Fisher LDA（8.40）或最佳评分（8.42）来得到稀疏判别问题。当 $p \gg N$ 并且特征没有结构（比如基因组分类问题）时，可设置 $\tilde{\boldsymbol{\Sigma}}_w$ 为对角矩阵 $\text{diag}(\hat{\boldsymbol{\Sigma}}_w)$，这时 Fisher LDA 算法很适合求解这类问题。因为此矩

阵是正定的，可让 $\boldsymbol{\Omega}=\mathbf{0}$，这样该问题很容易用针对惩罚矩阵分解的软阈值化算法来求解。当问题具有空间或时间结构时，所选择的 $\boldsymbol{\Omega}$ 要能使解在空间或时间上平滑。在这种情况下，最佳评分方法是很不错的选择，因为二次项可被吸收进二次损失函数中。否则，可将矩阵 $\boldsymbol{\Omega}$ 设为对角矩阵，下面的例子就是这样。此外，最佳评分很方便使用。R 包 penalizedLDA（Witten 2011）和 sparseLDA（Clemmensen 2012）分别对求解式（8.40）和式（8.42）的算法进行了实现。

示例：面部轮廓

下面通过 Clemmensen et al.（2011）给出的形态示例来阐述目标函数（8.42）的稀疏判别分析。数据集由 20 名成年男性和 19 名成年女性的面部轮廓构成。这里采用最小描述长度（Minimum Description Length，MDL）来标注轮廓，这源于 Thodberg and Olafsdottir（2003）的工作。接下来，在所得的 65 个 MDL 关键点（landmark）上采用 Procrustes 对齐。这 65 个关键点以坐标 (x,y) 的形式来表示，将其向量化后会得到 $p=130$ 个空间特征。在式（8.42）中，设 $\boldsymbol{\Omega}=\boldsymbol{I}$。在这种情况下，空间特征已经光滑，岭惩罚 $\boldsymbol{I}$ 足以处理强空间的自动相关。22 个面部轮廓（11 名女性，11 名男性）会作为训练数据，剩下 17 个面部轮廓（8 名女性，9 名男性）作为测试数据。图 8-8 的左图和中图是两类面部轮廓的示意图。

在训练数据上进行的留一（Leave-one-out）交叉验证，估计出了能够产生 10 个非零特征的最佳 λ 值。这里有两个类，因此只有一个稀疏方向。图 8-8 右图为非零权重示意图。模型中的少数关键点位于面部轮廓中高曲率点附近，这表明重要的性别差异都位于这些区域。训练和测试的正确分类率均为 82%。

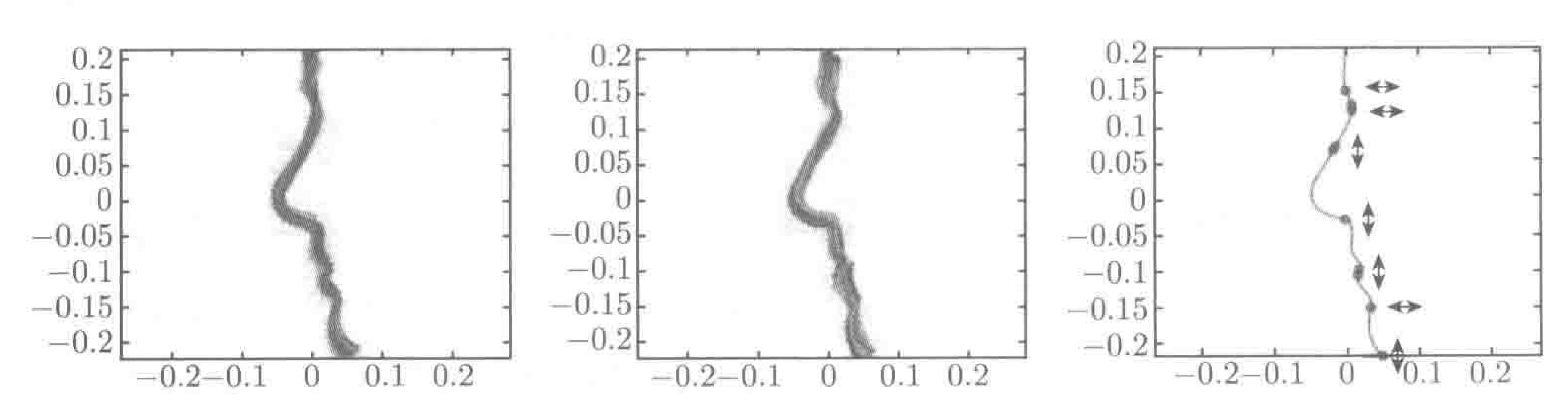


图 8-8 面部轮廓和 65 名女性（左图）及男性（中图）的 (x,y) 坐标。右图：面部轮廓的平均形状，SDA 模型中有 10 个坐标。叠加圆点表示保留在稀疏判别向量中的关键点。箭头表示男性和女性之间不同的方向（见彩插）

8.5 稀疏聚类

本节将讨论稀疏聚类方法，使用这些方法可以滤除不含有信息的特征。首先简

要给出聚类的背景。更多介绍参见文献 Hastie et al. (2009, Chapter 14)。

8.5.1 聚类的一些背景知识

假设对 N 个样本（有 p 个特征，$p \gg N$）进行聚类，目标是为具有 p 个特征的相似样本进行聚类。实现这种功能的标准方法称为“层次聚类”（hierarchical clustering）。更确切地说，这应称为自底向上（agglomerative，也称自顶向下）的层次聚类。此方法从单个的样本开始处理，然后根据某种度量来合并（或归组）最近的样本，通常选择欧几里得距离作为度量。继续此过程，在每个阶段将最接近的样本放在一起。在这种方式中，合并的不仅是单个样本对，前面步骤创建的聚簇（cluster）与单个样本或其他聚簇也会合并。因此，需要定义两个聚簇之间的相似程度（linkage measure），即两个聚簇之间的距离。一些常见的相似度包括：(1) average linkage，将两个簇中所有样本之间的距离加起，然后求平均；(2) complete linkage，两个簇中最远两个点的距离；(3) single linkage，两个簇中最近两个点的距离。

示例：有 6 个类的模拟数据

图 8-9 的上图是一个层次聚类的例子，数据是通过模拟产生的，总共有 120 个样本，每个样本有 2000 个特征。该图给出了基于欧几里得距离和基于 complete linkage 的层次聚类结果。聚类树（也称树状图，dendrogram）展示了合并的细节，最后在顶部得到单个簇。聚类时并不会用到叶子颜色，下面进行说明。

现在假设样本仅由一个特征子集构成。为了可解释性并提高聚类效果，可分离这个子集。图 8-9 的上图为生成的数据，即前 200 个特征的平均水平会在 6 个预定义类上变化，剩下的 1800 个特征为标准高斯噪声。这些类别不会在聚类中使用，但聚类完成后，会根据真实的分类情形来绘制树形图的彩色叶子。在此可以看到，层次聚类因无信息的特征而变得难以理解，并不能将样本进行很好地聚类。在这种情况下，最好分离出含有信息的特征子集，以此得到可解释性，并提高聚类效果。下面介绍这样的方法。

8.5.2 稀疏层次聚类

下面介绍针对上述问题引入稀疏性和特征选择的方法。设有一个数据矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，标准聚类的相似性度量 $D_{i,i'} = \sum_{j=1}^{p} d_{i,i',j}$(其中 $d_{i,i',j} = (x_{ij} - x_{i'j})^2$) 会采用欧几里得距离。这里要找到一组特征权重 $w_j \geqslant 0$，将权重相似度量定义为：$\tilde{D}_{i,i'} = \sum_{j=1}^{p} w_j d_{i,i',j}$，每个权重反映对应特征的重要性。最后，将这种修改过的相异（dissimilarity）矩阵作为层次聚类的输入。

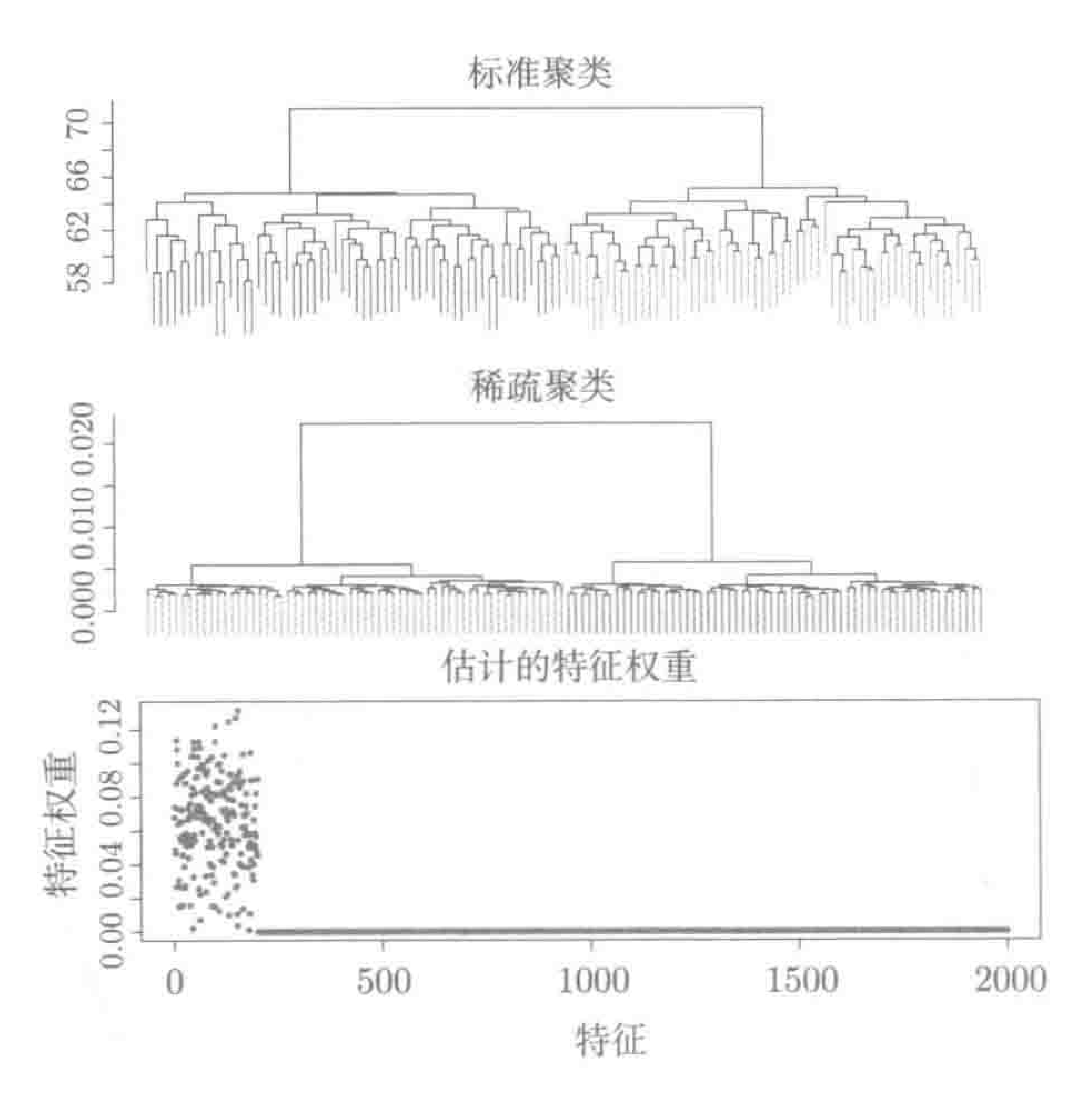


图 8-9　用在模拟数据上的标准聚类和稀疏聚类。所生的数据有 6 个类，这些类的前 200 个特征的平均水平不同，剩下的 1800 特征为相同的标准高斯噪声。上面两幅图分别为标准聚类和稀疏聚类的结果，所使用的相似度量为 complete linkage。每个样本所属的类会由叶子的颜色表示出来。下图显示了由稀疏聚类所估计的特征权重

$\boldsymbol{\Delta}$ 是一个 $N^2 \times p$ 的矩阵，它的第 j 列含有第 j 个特征的 N^2 对相异性。$\boldsymbol{\Delta}_1$ 是 $D_{i,i'}$ 的向量化，$\boldsymbol{\Delta}w$ 是 $\tilde{D}_{i,i'}$ 的向量化。现在需要找到一个稀疏且受一些归一化限制的向量 w，使其能涵盖 $\boldsymbol{\Delta}$ 的大部分变化。这会引出一个惩罚矩阵分解问题（Witten et al. 2009，见 7.6 节）

$$\underset{\boldsymbol{u}\in\mathbb{R}^{N^2}, w\in\mathbb{R}^p}{\text{maximize}} \left\{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Delta}w\right\}$$
$$\text{其约束为 } \|\boldsymbol{u}\|_2 \leqslant 1,\ \|w\|_2 \leqslant 1, \|w\|_1 \leqslant s,\ \text{且} w \succeq 0 \tag{8.43}$$

注意，w_j 是第 j 个特征的相异矩阵的权重。假设得到优化解 $\tilde{w}$，可将 $\boldsymbol{\Delta}w$ 中的元素重排成 $N \times N$ 矩阵，并在重新加权后的相异矩阵上执行层次聚类。这就会得到数据的稀疏层次聚类。图 8-9 所示为稀疏聚类分离出的含有信息的特征（下图），使用这样的信息可正确地将样本聚集到预定义的类中（中图）。

8.5.3　稀疏 K 均值聚类

K 均值是另一种常用的聚类方法。该方法需要预定义 K 个类，然后尝试将样本分到 K 个不同类中。每个类都有一个中心点，样本离哪个类的中心最近，就会分配给那个类。

具体而言，K 均值算法会维持索引集 $\{1,2,\ldots N\}$ 的一个分区 $\mathcal{C}=\{C_1,\ldots,C_k\}$，其中 C_k 表示分配给第 k 个类的样本索引。最小化目标函数

$$W(\mathcal{C})=\sum_{k=1}^{K}\sum_{i\in C_k}\|x_i-\bar{x}_k\|_2^2 \tag{8.44}$$

可以得到这些分区，这里的 x_i 表示第 i 个样本，$\bar{x}_k$ 是一个 p 维向量，表示第 k 类样本的均值。集合 $\{\bar{x}_k\}_1^K$ 称为编码本（codebook），编码器 $\tau(i)$ 将每个样本 x_i 分配给最靠近的某个聚类 k，因此 $C_k=\{i:\tau(i)=k\}$。标准 K 均值聚类算法会交替优化 $\mathcal{C}$ 和 $\{\bar{x}_1,\ldots,\bar{x}_K\}$，并找到 $W(\mathcal{C})$ 的局部最优解。标准 K 均值聚类算法的目标函数为

$$\sum_{i,i'\in C_k}\|x_i-x_{i'}\|_2^2=2N_k\sum_{i\in C_k}\|x_i-\bar{x}_k\|_2^2 \tag{8.45}$$

其中 $N_k=|C_k|$，可通过平方欧几里得相异矩阵 $D_{i,i'}$ 来交替得到 K 均值聚类。对于一般的相异矩阵而言，K-medoids 聚类是一个自然的推广（Hastie et al. 2009）。

一种定义稀疏 K 均值聚类目标函数的合理方式，是使用最小化加权聚簇中的平方和

$$\underset{\mathcal{C},w\in\mathbb{R}^p}{\text{minimize}}\left\{\sum_{j=1}^{p}w_j\left(\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i,i'\in C_k}d_{i,i',j}\right)\right\}$$

这里还需要对 w 增加约束，使问题有意义。增加的约束为 $\|w\|_2\leqslant 1$，$\|w\|_1\leqslant s$，另外还要增加一个非负约束 $w\succeq 0$，使该问题成为关于 w 的凸函数，但会导致病态解 $\hat{w}=0$。另一方面，采用约束 $\|w\|_2\geqslant 1$，$\|w\|_1\geqslant s$，$w\succeq 0$ 会得到有用的解，但会得到一个关于 w 的非凸问题。

Witten and Tibshirani（2010）提出了一种改进的方法，关注聚簇间的平方和

$$\underset{\mathcal{C},w\in\mathbb{R}^p}{\text{maximize}}\left\{\sum_{j=1}^{p}w_j\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{i'=1}^{N}d_{i,i',j}-\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i,i'\in C_k}d_{i,i',j}\right)\right\}$$
$$\text{其约束为 } \|w\|_2\leqslant 1,\|w\|_1\leqslant s,w\succeq 0 \tag{8.46}$$

对于所有 j，有 $w_j=1$，则由式（8.45）可知，式（8.46）的第二项等于 $2W(\mathcal{C})$。因此，该方法等价于 K 均值。这是关于 w 的凸问题，通常能得到全局最优解，可用简单的交替迭代算法来求解。固定 $\mathcal{C}=\{C_1,\ldots,C_k\}$，最小化一个关于 w 的凸问题，这可通过软阈值来求解。固定 w，优化关于 $\mathcal{C}$ 的目标函数，这其实是一个权重 K 均值算法。整个算法的详细求解过程可以参见习题 8.11。

8.5.4 凸聚类

K 均值聚类及其稀疏形式是双凸问题，因此难以保证全局解。下面介绍另外一种聚类方法。这个方法能够得到一个凸规划问题，可用其来替代 K 均值和层次

聚类。前面的方法都用稀疏来做特征选择，该方法则是使用稀疏来确定聚类数量和每个类中的样本。

假设有 N 个样本，为第 i 个样本 $x_i \in \mathbb{R}^p$ 分配一个厚型 $u_i \in \mathbb{R}^p$。然后最小化目标函数

$$J(u_1, u_2, \ldots, u_N) = \frac{1}{2}\sum_{i=1}^{N} \|x_i - u_i\|^2 + \lambda \sum_{i<i'} w_{ii'} \|u_i - u_i'\|_q \tag{8.47}$$

对于给定的 $\lambda \geqslant 0$ 及某个 q 范数（通常取 $q=1$ 或 $q=2$），要找到的 u_i 应尽量与相应的 x_i 接近，但 u_i 之间不能离得太远。权重 $w_{i,i'}$ 可设为 1，或将其看成样本 i 与 i' 之间距离的函数。注意，若 $q \geqslant 1$，则该问题是凸的，通常选择 $q=2$（组 lasso）。惩罚项会让 u_i 对应的向量中各个元素彼此靠近，若 λ 的值足够大，有些元素之间的距离可能会为 0。

在得到的解中，每一个不同的 $\hat{u}_i$ 都表示一个聚类。但不要认为 $\hat{u}_i$ 就是某个聚类的中心（如图 8-10 所示）。这个例子有两个类，它们各有 50 个球形高斯数据点，它们的均值在这两个方向上都相差三个单位。这里的 $q=2$，权重函数为 $w_{ii'} = \exp(-\|x_i - x_{i'}\|^2)$。在图 8-10 右图中，聚类的颜色与真实情况一致。该图的标题会进一步说明这一点。这种方法很有吸引力，因为目标函数是凸的，而且它能选择聚类数，也可以得到含有丰富信息的特征。

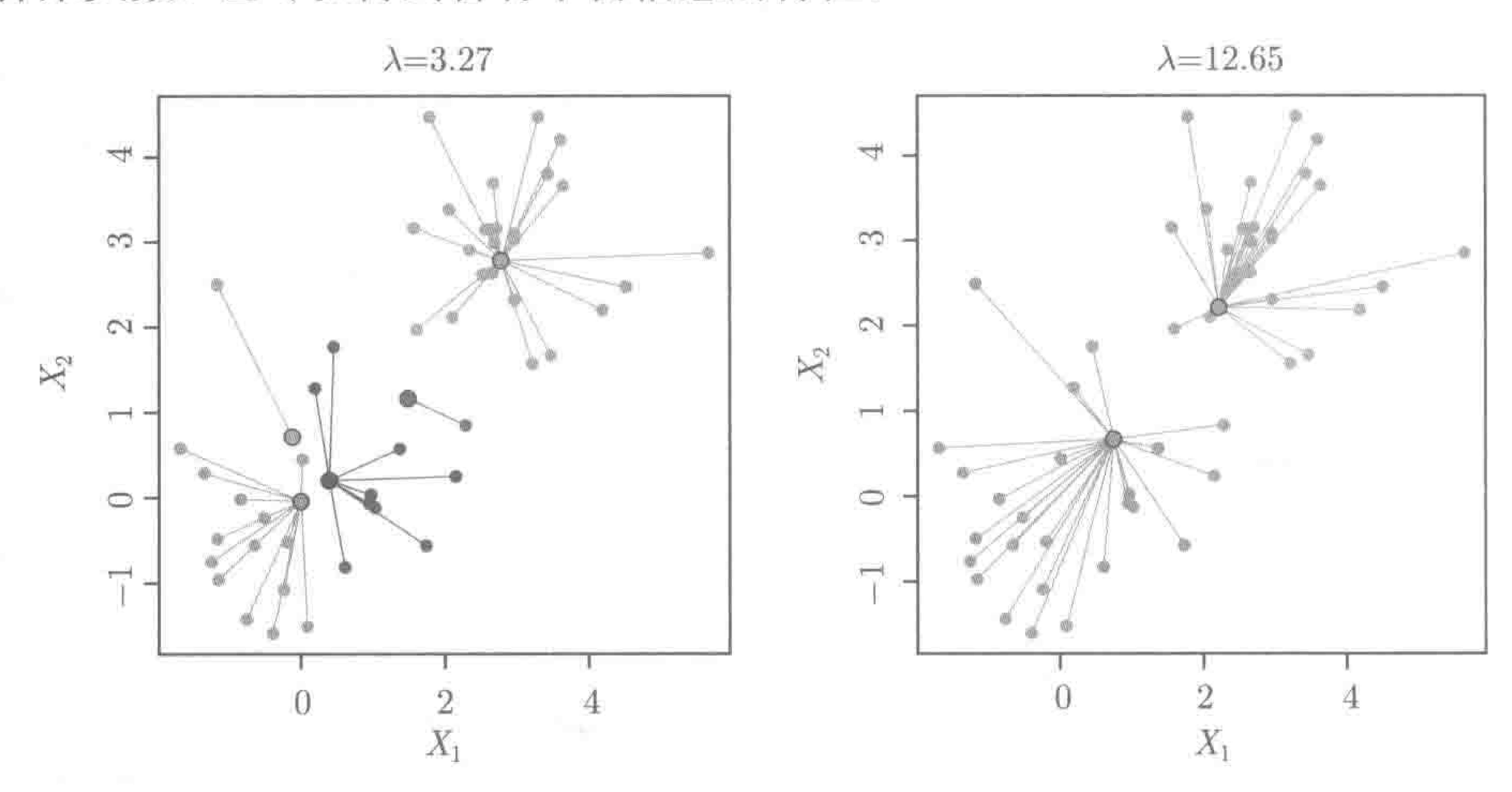


图 8-10 在生成的数据上使用凸聚类，这些数据来自两个球形高斯总群（population），这些总群在每个方向被分开。这里给出 50 个 λ 值的解路径。右图给出得到两个聚类的最小 λ 值，在这种情况下，真实簇识别了出来。点 x_i 与颜色相同的原型 $\hat{\mu}_i$ 关联。估计原型不必接近聚类的中心

接下来的例子来自 Chi and Lange（2014），根据齿系对哺乳动物聚类。总共有 27 类哺乳动物，每类有 8 种不同类型的牙齿：顶门齿、底部门齿、顶部犬齿、底部

犬齿、前臼齿顶部、底部前磨牙、顶级臼齿和底部臼齿。图 8-11 给出了不同的 λ 取值所得到的聚类结果。目标函数中的权重 w_{ii} 采用的是基于核的权重。为了可视化，原型[①]投影到前两个主成分上。连续求解路径展示了哺乳动物中的相似性。这些例子都采用了 R 的 cvxcluster 包（Chi and Lange 2014）。解的路径生成一棵树，它在本例中相当于基于 average linkage 的层次聚类。

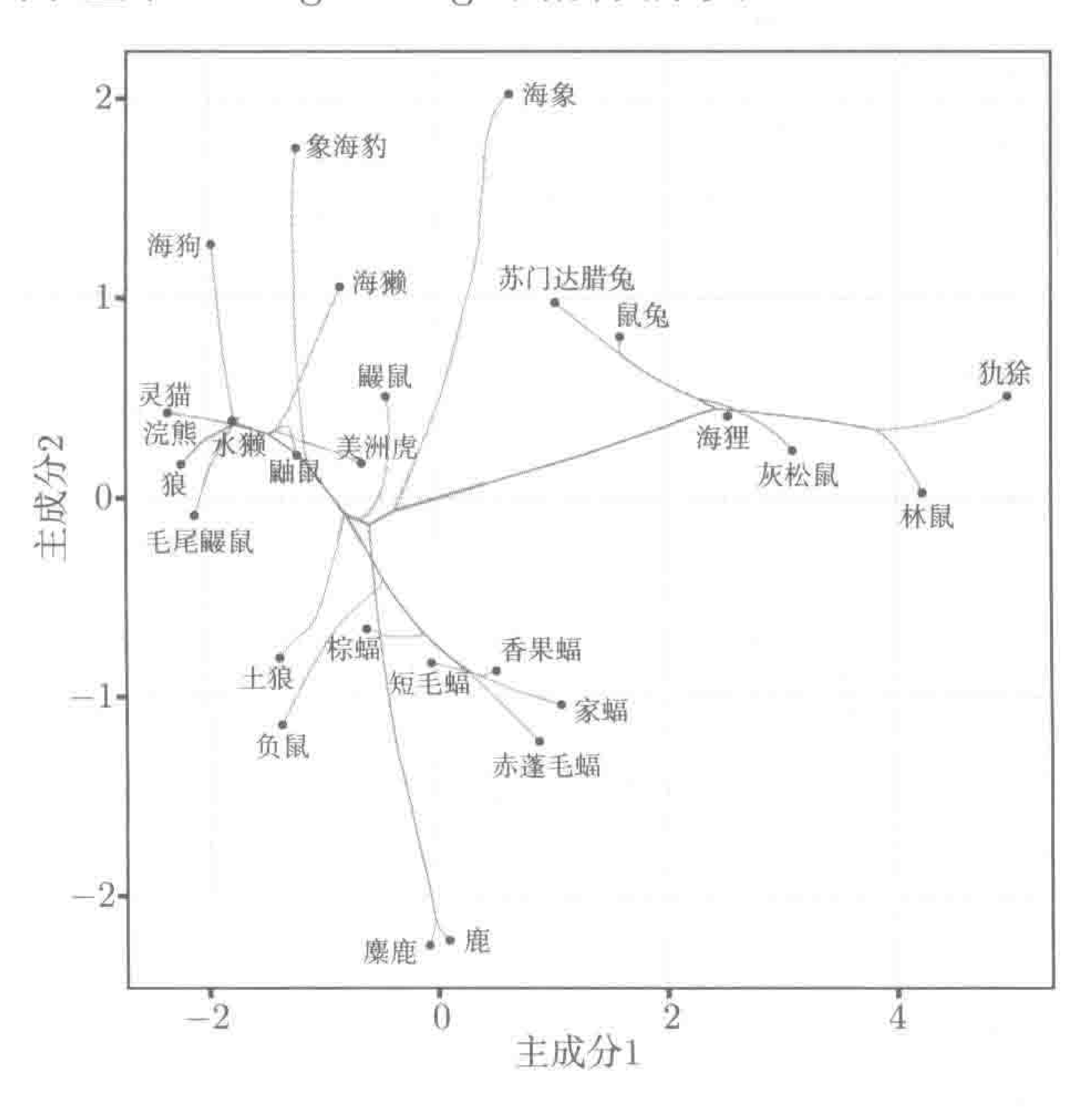


图 8-11 哺乳动物数据：凸聚类 [见式（8.47）] 解的路径，λ 值的网格。当 λ 增加时，原型会集中到一个较小集合中

参考文献注释

Jolliffe et al.（2003）针对稀疏 PCA 提出了 SCoTLASS 方法（8.7）。Witten et al.（2009）重写了式（8.9），并用交替迭代来求解。d'Aspremont et al.（2007）对非凸的 SCoTLASS 采用半定规划松弛。Amini and Wainwright（2009）就稀疏 PCA 对这种松弛的特征选择性质给出了一些理论结果。Zou et al.（2006）提出了基于式（8.13）的重构误差形式。Johnstone（2001）研究了普通 PCA 的高维渐近性，并提出了尖协方差模型（8.22）。Johnstone and Lu（2009）和 Birnbaum et al.（2013）研究了用于估计稀疏主成分的各种两阶段算法。当特征向量属于 ℓ_q 球时，Birnbaum et al.（2013）和 Vu and Lei（2012）通过 ℓ_2 的误差估计来推导稀疏 PCA 的极小极

① 此处原型即 $\boldsymbol{u}_i$。—— 译者注

大（minimax）下界。Ma（2010，2013）和 Yuan and Zhang（2013）将 power 方法与软阈值结合起来研究稀疏 PCA 的迭代算法。Berthet and Rigollet（2013）研究了高维稀疏 PCA 的检测问题，并得到了相关计算的复杂度，这与随机 K 团（k-clique）问题有关。

Olshausen and Field（1996）提出了 8.2.5 节所讨论的稀疏编码。大量深度学习的文献包括 Le et al.（2012）以及近期的一些文献。

研究稀疏典型相关分析的文章有很多，包括：Parkhomenko,Tritchler and Beyene（2009）、Waaijenborg,Vers'elewel de Witt Hamer and Zwinderman（2008）、Phardoon and Shawe-Taylor（2009）、Parkhomenko et al.（2009）、Witten et al.（2009）、Witten and Tibshirani（2009），以及 Lykou and Whittaker（2010）。Dudoit，Fridlyand and Speed（2002）基于芯片数据对不同的分类方法（包括对角线 LDA）进行比较。Tibshirani,Hastie,Narasimhan and Chu（2001）和 Tibshirani（2003）提出了最近收缩中心算法。Trendafilov and Jolliffe（2007）、Leng（2008）、Clemmensen et al.（2011）以及 Witten and Tibshirani（2011）研究了稀疏判别分析，包括 Fisher 的框架和最佳评分。Lange，Hunter and Yang（2000）、Lange（2004）以及 Hunter and Lange（2004）研究了 Minorization 算法。稀疏层次聚类和稀疏 K 均值聚类由 Witten and Tibshirani（2010）提出，而凸聚类由 Pelckmans, De Moor and Suykens（2005）和 Hocking, Vert, Bach and Joulin（2011）提出。

习　　题

习题 8.1　这道习题主要针对主成分分析的一些基本特性。

(a) 求证：第一主成分是样本协方差矩阵 $\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的最大特征向量。

(b) 假设 $\boldsymbol{X}$ 的行 $\{x_1,\ldots,x_N\}$ 独立同分布，其中分布 $\mathbb{P}$ 的均值为零，并假设 $\boldsymbol{\Sigma}=\mathrm{Cov}(x)$ 有唯一的最大特征值 λ。解释为什么对于较大的样本数量 N，可能的期望 $\hat{v}$ 会接近

$$v^*=\underset{\|v\|_2=1}{\arg\max}\ \mathrm{Var}(v^{\mathrm{T}}x)，其中\ x\sim\mathbb{P} \tag{8.48}$$

习题 8.2　式（8.1）的主成分可分别定义为 v_1、v_2，等等。

(a) 求证：主成分 $\boldsymbol{z}_j$ 互不相关。

(b) 求证：可使用 $\boldsymbol{z}_j$ 的不相关性，而不使用 v_j 的正交性来定义主成分方向的序列。

习题 8.3　设 $\mu=0$, 对于重构误差目标函数（8.4）。

(a) 固定 $\boldsymbol{V}_r$，求证：重建权重的最优选择为 $\lambda_i=\boldsymbol{V}_r^{\mathrm{T}}x_i$，$i=1,\ldots,N$。

(b) 使用 (a) 的结果证明：最优解 $\boldsymbol{V}_r$ 会让 $\boldsymbol{V}_r^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}$ 最大化，并可通过截断 SVD 来得到。

习题 8.4　考虑式（8.8）和用来求解的算法 8.1。部分最大化 $\boldsymbol{u}$（其中 $\|\boldsymbol{u}\|_2 \leqslant 1$），求证任何关于 v 的稳定值也是 SCoTLASS 方法（8.7）的稳定值。

习题 8.5　考虑问题

$$\underset{\boldsymbol{u},v,d}{\text{minimize}} \left\|\boldsymbol{X} - d\boldsymbol{u}v^{\mathrm{T}}\right\|_F$$
$$\text{其约束为 } \|\boldsymbol{u}\|_1 \leqslant c_1，\|v\|_1 \leqslant c_2，\|\boldsymbol{u}\|_2 = 1，\|v\|_2 = 1，d \geqslant 0 \tag{8.49}$$

其中，$\boldsymbol{X}$ 为 $N \times p$ 矩阵。假设 $1 \leqslant c_1 \leqslant \sqrt{N}$，$1 \leqslant c_2 \leqslant \sqrt{p}$，求证式（8.49）的一个解 $\boldsymbol{u}_1$、v_1、d_1 也是下面优化问题的解：

$$\underset{u,v}{\text{maximize}}\ \ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}v$$
$$\text{其约束为 } \|\boldsymbol{u}\|_1 \leqslant c_1，\|v\|_1 \leqslant c_2，\|\boldsymbol{u}\|_2 = 1，$$
$$\|v\|_2 = 1，d_1 = \boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{X}v_1 \tag{8.50}$$

习题 8.6　本习题关注 SCoTLASS 方法（8.7）的特性。

(a) 用 Cauchy–Schwarz 不等式证明算法 8.1 的不动点是 SCoTLASS 的局部最小值。

(b) 由 Cauchy–Schwarz 不等式可以得到

$$v^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v \geqslant \frac{(v^{(m)\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v)^2}{v^{(m)\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v^{(m)}} \tag{8.51}$$

当 $v = v^{(m)}$ 时，等号成立，所以在 $v^{(m)}$ 处，$\frac{\left(v^{(m)\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v\right)^2}{v^{(m)\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}v^{(m)}}$ 是 $v^{\mathrm{T}}\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X}v$ 的最小值。求证：使用这个 minorization 函数，可由 MM 算法（见 5.8 节）得到算法 8.1。

习题 8.7　求证：问题（8.15）在增加约束 $\|\theta\|_1 \leqslant t$ 后得到的解也是 SCoTLASS 问题（8.7）的解。

习题 8.8　考虑问题：

$$\underset{\theta:\|\theta\|_2=1}{\text{minimize}} \sum_{i=1}^{N} \|x_i - \theta z_i\|_2^2 \tag{8.52}$$

其中，$\{x_i\}_{i=1}^N$ 和 θ 都是 p 维向量，变量 $\{z_i\}_{i=1}^N$ 为标量。求证：这个问题的最优解是唯一的，且为 $\hat{\theta} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}}{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}\|_2}$。

习题 8.9　求证：式（8.17）中的向量 $\boldsymbol{u}_k$ 可通过求解多因子稀疏 PCA 问题（8.16）得到。

习题 8.10 关于稀疏主成分重建问题（8.13），解答

(a) 固定 $\boldsymbol{V}$，求解 $\boldsymbol{\Theta}$。

(b) 求证：对于第 k 次迭代，当 $\lambda_{1k}=0$ 时，关于 $\boldsymbol{X}$ 的任意 k 个主成分的集合是稳定的；具体而言，若该算法从最大 k 个主成分开始，它就不会再执行下去了。

(c) 求证：在条件 (b) 下，$\boldsymbol{\Theta}=\boldsymbol{V}$ 可让该目标函数最大化，并且都等于 $\boldsymbol{V}_k$，这个矩阵由 $\boldsymbol{X}$ 的最大 k 个主成分构成。

(d) 对于 (c) 中的解 $\boldsymbol{V}_k$，求证 $\boldsymbol{V}_k\boldsymbol{R}$ 也是一个解，其中 $\boldsymbol{R}$ 为 $k\times k$ 正交矩阵。

因此，该版本的稀疏主成分类似于采用基于旋转因子的最大方差法（Kaiser 1958）所实现的稀疏性。

习题 8.11 本习题关注稀疏的 K 均值聚类算法。关于目标函数（8.46），解答

(a) 证明：固定 w，优化问题

$$\underset{C_1,\ldots,C_K}{\text{minimize}}\left\{\sum_{j=1}^{p}w_j\left(\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i,i'\in C_k}(x_{ij}-x_{i'j})^2\right)\right\} \tag{8.53}$$

其中 $\mathcal{C}=(C_1,\ldots,C_k)$。这可以看作加权数据的 K 均值聚类。给出这个问题的求解算法。

(b) 固定 $\mathcal{C}=(C_1,\ldots,C_k)$，对 w 进行优化

$$\underset{w\in\mathbb{R}^p}{\text{maximize}}\left\{\sum_{j=1}^{p}w_j\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{i'=1}^{N}(x_{ij}-x_{i'j})^2-\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i,i'\in C_k}(x_{ij}-x_{i'j})^2\right)\right\}$$

其约束为 $\|w\|_2=1$，$\|w\|_1\leqslant s$，且$w_j\geqslant 0$ (8.54)

这个问题与简单的凸问题

$$\underset{w\in\mathbb{R}^p}{\text{maximize}}\{w^{\mathrm{T}}a\}\ \text{，其中}\ \|w\|_2=1,\ \|w\|_1\leqslant s,\ \text{且}w\succeq 0 \tag{8.55}$$

形式相同。求证此问题并给出算法。

习题 8.12 有如下优化问题：

$$\underset{\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{p\times m}}{\text{minimize}}\left\{\sum_{i=1}^{N}\left\|x_i-\boldsymbol{A}\boldsymbol{B}^T x_i\right\|^2\right\} \tag{8.56}$$

其中，$\boldsymbol{x}_i\in\mathbb{R}^p, i=1,\ldots,N$ 是 $\boldsymbol{X}$ 的行，$m<\min(N,p)$。求证：这个问题的解满足 $\hat{\boldsymbol{A}}\hat{\boldsymbol{B}}=\boldsymbol{V}_m\boldsymbol{V}_m^{\mathrm{T}}$，其中，$\boldsymbol{V}_m$ 是矩阵 $\boldsymbol{X}$ 的前 m 个右奇异向量。

习题 8.13　思考优化问题（8.21）中给定的稀疏编码器，给出一种简单的交替算法来求解该问题，阐明每一步的实现细节。

习题 8.14　交替回归的典型相关：考虑两个随机向量 $\boldsymbol{X} \in \mathbb{R}^{m_1}$和$\boldsymbol{Y} \in \mathbb{R}^{m_2}$，它们的协方差矩阵分别为 $\boldsymbol{\Sigma}_{11}$和$\boldsymbol{\Sigma}_{22}$，交叉协方差矩阵为 $\boldsymbol{\Sigma}_{12}$。设 $\boldsymbol{L} = \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$，用 γ_i和τ_i 表示 $\boldsymbol{L}$ 的左奇异向量和右奇异向量，与之对应的奇异值为 ρ_i。

(a) 求证：β_1 和 θ_1 可让 $\mathrm{Corr}(X\beta, Y\theta)$ 最大化，则

$$\beta_1 = \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\gamma_1，\theta_1 = \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\tau_1 \tag{8.57}$$

成立，并且最大相关为 ρ_1。

(b) 下面来研究类似问题。数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的大小分别为 $N \times p$ 和 $Y \times p$，每个列的均值为 0。在这种情形下，以它们的样本估计来替换掉 $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_{12}$，则可得到典型相关估计。在此基础上，求证：最优样本典型向量为 $\beta_1 = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-\frac{1}{2}}\boldsymbol{\gamma}_1$和$\theta_1 = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-\frac{1}{2}}\boldsymbol{\gamma}_1$，其中 $\boldsymbol{\gamma}_1$和$\boldsymbol{\tau}_1$ 是矩阵 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-\frac{1}{2}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-\frac{1}{2}}$ 的最大左奇异向量和右奇异向量。

(c) 设第一相关变量分别为 $\boldsymbol{z}_1 = \boldsymbol{X}\beta_1$和$\boldsymbol{s}_1 = \boldsymbol{Y}\theta_1$，它们的维数都为 N，设 $\boldsymbol{H}_X = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ 投影到 $\boldsymbol{X}$ 的列空间上，$\boldsymbol{H}_Y$ 也与之类似。求证：

$$\boldsymbol{H}_{\boldsymbol{X}}\boldsymbol{s}_1 = \rho_1\boldsymbol{z}_1，\boldsymbol{H}_{\boldsymbol{Y}}\boldsymbol{z}_1 = \rho_1\boldsymbol{s}_1$$

之后，交替回归 $\boldsymbol{X}$和$\boldsymbol{Y}$，直到收敛到最大典型相关问题的解。

习题 8.15　由习题 8.14 可知，交替最小二乘回归可得到一对最大典型变量。如何通过修改这种典型变量的求解方法来得到第二大的典型变量 $(\boldsymbol{z}_2, \boldsymbol{s}_2)$? 做出修改并证明相关算法正确。阐明如何扩展这种方法以找到所有后续结果。

习题 8.16　本习题关注基于最佳评分的 CCA。给定的数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 与 8.14 一样，它们列的均值为 0，并且都是列满秩，考虑问题

$$\underset{\beta,\theta}{\mathrm{minimize}} \|\boldsymbol{Y}\theta - \boldsymbol{X}\beta\|_2^2 \quad \text{，其约束为} \ \frac{1}{N}\|\boldsymbol{Y}\theta\|_2 = 1 \tag{8.58}$$

(a) 固定 θ 求解 β，然后固定 β，求解 θ，从而得到这个问题的解。给出详细的实现过程。

(b) 求证：原问题的最优解为 $\hat{\theta} = \theta_1$，$\hat{\beta} = \rho_1\beta_1$，其中 θ_1和β_1 为第一对典型向量，ρ_1 为最大的典型相关。

(c) 求证：除了有 $\|\boldsymbol{Y}\theta_1 - \boldsymbol{X}\beta_1\|_2^2 = 1 - \rho_1^2$，这个等式还表明，求解最佳评分问题等价于求解 CCA 问题。

(d) 描述如何找到后续解，这些解要与前面的解不相关。给出如何通过数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的变换得到这样的后续解。

(e) 如果有约束 $\|\boldsymbol{X}\beta\|_2 = 1$，则求解该算法是否会变化？

习题 8.17 本习题关注低秩 CCA。假设习题 8.14 和习题 8.16 中的矩阵至少有一个不是列满秩（比如，$N < \min(p, q)$），解答

(a) 求证：$\rho_1 = 1$ 时，CCA 问题有多个最优解。

(b) 假设 $\boldsymbol{Y}$ 是列满秩，但 $\boldsymbol{X}$ 不是，若为式（8.58）增加岭约束，即求解优化问题

$$\underset{\beta,\theta}{\text{minimize}} \|\boldsymbol{Y}\theta - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2\text{，其约束为 } \|\boldsymbol{Y}\theta\|_2 = 1 \tag{8.59}$$

这个问题会退化吗？描述这个解。

(c) 求证：式（8.59）的解等价于将 CCA 用到了 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 上，不同之处在于 $\hat{\beta}$ 要满足归一化条件 $\frac{1}{N}\beta^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}\right)\hat{\beta} = 1$，通过岭化协方差估计来得到相应的归一化。

习题 8.18 数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的列都进行了中心化，考虑优化问题

$$\underset{\beta,\theta}{\text{maximize}} \left\{\widehat{\text{Cov}}(\boldsymbol{X}\beta, \boldsymbol{Y}\theta) - \lambda_1 \|\beta\|_2^2 - \lambda_2 \|\beta\|_2^2 - \lambda_1' \|\theta\|_1 - \lambda_2' \|\theta\|_2\right\} \tag{8.60}$$

使用习题 8.14 的结果，给出如何使用交替弹性网拟和方法而不是最小二乘回归来求解这个问题。

习题 8.19 本习题关注稀疏典型相关分析。对习题 8.16 中的最佳评分问题（8.58）增加一个 ℓ_1 约束

$$\underset{\beta,\theta}{\text{minimize}} \left\{\|\boldsymbol{Y}\theta - \boldsymbol{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\theta\|_1\right\}\text{，其约束为 } \|\boldsymbol{X}\beta\|_2 = 1 \tag{8.61}$$

使用习题 8.14、习题 8.16 和习题 8.17 的结果，并采用交替弹性网拟和方法而不是最小二乘回归来求解这个问题。

习题 8.20 对于 8.4.1 节的多元高斯情形，为每个类假设不同的协方差矩阵 $\boldsymbol{\Sigma}_k$，并解答

(a) 求证：判别函数 δ_k 是 x 的二次函数。并给出决策界。

(b) 假设协方差矩阵 $\boldsymbol{\Sigma}_k$ 是对角矩阵，这意味着每个类的特征 $\boldsymbol{X}$ 条件独立。用朴素贝叶斯分类器来描述第 k 类和第 ℓ 类之间的决策界。

习题 8.21 本习题与 8.4.2 节的最近收缩中心问题有关。思考有 ℓ_1 正则化的目标函数

$$\underset{\substack{\bar{\mu}\in R^p \\ \alpha_k\in\mathbb{R}^p, k=1,\ldots,p}}{\text{minimize}} \left\{\frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j=1}^{p}\frac{(x_{ij} - \bar{\mu}_j - \alpha_{jk})^2}{s_j^2} + \lambda\sum_{k=1}^{K}\sum_{j=1}^{p}\frac{\sqrt{N_k}}{s_j}|\alpha_{jk}|\right\} \tag{8.62}$$

这里将每个类的均值分解成总体均值加上每个类与总体均值之差，将类的大小和每个特征的类内标准偏差作为加权惩罚项。

(a) 求证：用第 j 个特征的总体均值 $\bar{x}_j$ 替换 $\bar{\mu}_j$，会得到式 (8.35a) 那样的收缩方案，只是在 m_k 中没有 $1/N$ 这一项。

(b) 求证：如果不像上面那样限制 $\bar{\mu}_j$，则 (a) 得不到式 (8.33) 那样的解。

(c) 对于 $j = 1, \dots, p$，增加约束 $\sum_{k=1}^{K} \alpha_{jk} = 0$。讨论这个问题的解，它是否与 (a) 的解一致？

习题 8.22　求证：基于惩罚的 Fisher 判别问题 (8.40) 和基于惩罚的最佳得分问题 (8.42) 存在这样的等价关系：一个问题的固定点也是另一个问题的固定点。(Clemmensen et al. 2011 和 Witten and Tibshirani 2011。)

第 9 章　图和模型选择

9.1　引　　言

概率图模型为建立高维数据的简约模型提供了一种实用的框架，其基础是概率论和图论的交集，其中图的特性指定了一组随机变量的条件独立特性。在典型应用中，图的结构是未知的，重点在于从样本估计图的结构，这就是图模型选择问题。本章从这个目的出发，讨论了一系列基于 ℓ_1 正则化的方法。

9.2　图模型基础

在此先简单介绍一下图模型的基础，更多细节参见本章最后的文献注释。任意随机变量集合 $X=(X_1,X_2,\ldots,X_p)$ 都可以和一些潜在图模型的顶点集合 $V=\{1,2,\ldots,p\}$ 相关联。图模型的基本思想是用潜在图的结构（团结构或者割集结构）来限制随机向量 X 的分布。下面精讲几个概念。

9.2.1　分解和马尔可夫特性

一个普通图 G 包含顶点集 $V=\{1,2,\ldots,p\}$ 和边集 $E\subset V\times V$。本章仅关注无向图模型，即在边 $(s,\ t)\in E$ 和边 $(t,\ s)$ 之间没有方向。与之相比，**有向无环图**（Directed Acyclic Graph，DAG）更为流行，其中边具有方向。一般来说，这种有向图比无向图更加难以处理，这里并不关注。但是我们注意到，一些无向图的计算方法在 DAG 情况下也可用，请参考文献注释。

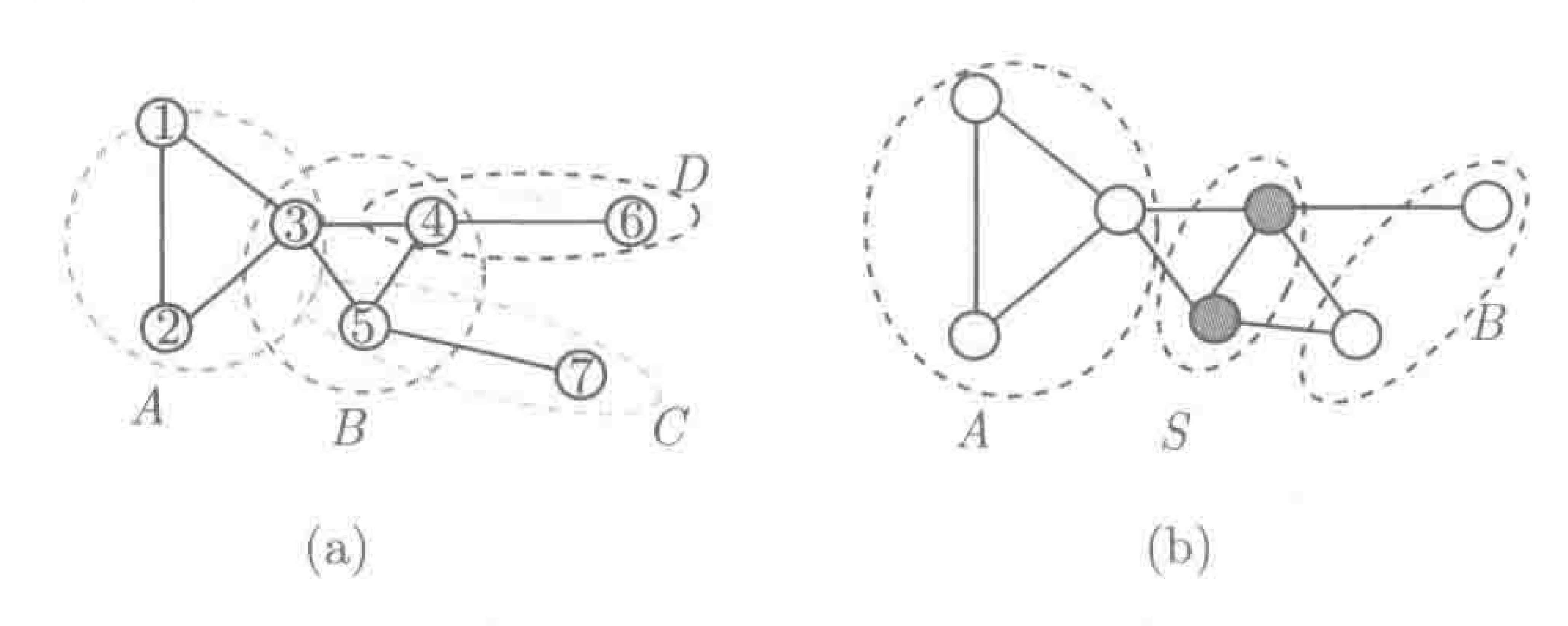


图 9-1　(a) 图中团的示意图，四个子集为团。集合 A 和 B 是 3 团，C 和 D 是 2 团，这通常称为边。所有这些团都是最大的。(b) 顶点割集 S 示意图，移除 S 中的顶点后，图分割为两个部分 A 和 B

一个**图团**（graph clique）$C \subseteq V$ 是顶点集的完全连接子集，即对所有的 $s, t \in C$，有 $(s, t) \in E$。如果一个团没有严格包含于其他团中，则称该团为最大团。例如，任意一个单顶点$\{s\}$自身是一个团，但是除非 s 为独立顶点（意味着它并不包含于任何边中），否则它不是最大团。$\mathfrak{C}$ 表示图中的所有团集，包括最大和非最大的。详见图 9.1a 对图团的解释。

1. 分解性

这里我们描述如何用图的团结构来限制随机向量 $(X_1 \ldots X_p)$ 的概率分布，其中随机向量按图中顶点排列。对于一个给定团 $C \in \mathfrak{C}$，**调和函数**ψ_C 是子向量 $x_C := (x_s, s \in C)$ 的实数值函数，为正实数值。给定这样一个调和函数集，如果概率分布 $\mathbb{P}$ 有分解形式为

$$\mathbb{P}(x_1, \ldots, x_p) = \frac{1}{Z} \prod_{C \in \mathfrak{C}} \psi_C(x_C) \tag{9.1}$$

则我们就说它**因式分解**G。这里的 Z 即是配分函数，$Z = \sum_{x \in \mathcal{X}^p} \prod_{C \in \mathfrak{C}} \psi_C(x_C)$。这就保证 $\mathbb{P}$ 会适当地归一化，并且定义了一个有效的概率分布。作为一个特例，任意概率分布对图 9-1a 因式分解必定有形式

$$\mathbb{P}(x_1, \ldots, x_7) = \frac{1}{Z} \psi_{123}(x_1, x_2, x_3) \psi_{345}(x_3, x_4, x_5) \psi_{46}(x_4, x_6) \psi_{57}(x_5, x_7) \tag{9.2}$$

这里涉及一些调和函数 $\{\psi_{123}, \psi_{345}, \psi_{46}, \psi_{57}\}$ 的选择。

式（9.1）的因式分解实际上非常重要，因为在团维度都不是特别大的情况下，这可以在存储和计算上带来显著的简化。例如，如果各个变量 X_s 都是二值的，则向量 $X \in \{-1, +1\}^p$ 上的概率分布需指定 $2^p - 1$ 个非负数，因而随图的大小呈指数级增长。另一方面，对一个基于团的因式分解，自由度最大为 $|\mathfrak{C}|\, 2^c$，其中 c 是任意团的最大基数。因此，针对基于团的因式分解，复杂度随着最大团维度 c 呈指数级增长，但是随着团数 $|\mathfrak{C}|$ 呈线性增长。幸运的是，许多我们感兴趣的实用模型可以指定为有界团的形式，在这种形式下基于团的表达式会十分有效用。

2. 马尔可夫特性

我们现在换一种方式思考问题，图结构可以基于其割集（见图 9-1b）用于约束 X 的分布。特别是，考虑割集 S 将图分割为不相连的 A 和 B，这里我们引入符号$\perp\!\!\!\perp$表示**条件独立于**。通过这个符号，我们可以说，如果

$$X_A \perp\!\!\!\perp X_B | X_S, \quad S \subset V \tag{9.3}$$

则随机向量 X 对 G 马尔可夫。图 9-1b 所示即是这种条件独立关系的一个例子。

马尔可夫链是该特性的一个特例。自然，它基于一种链结构图，有边集

$$E = \{(1,2), (2,3), \ldots, (p-1, p)\}$$

在该图中，任意单个顶点 $s \in \{2, 3, \ldots, p\text{-}1\}$可形成一个割集，将图分割为前部 $P=\{1, \ldots, s\text{-}1\}$和后部 $F=\{s+1, \ldots, p\}$。对这些割集，马尔可夫特性（9.3）得出的信息为：对马尔可夫链，给定当前 X_s，后部 X_F 条件独立于前部 X_P。当然，更多结构的图对应更复杂的割集，因此会有更多有趣的条件独立特性。

3. 分解性和马尔可夫特性的等价性

Hammersley–Clifford 定理是一个十分精辟的真理，即对任意严格为正的分布（对所有 $x \in \chi^p$，$\mathbb{P}(x) > 0$），这两个特性是等价的：当且仅当随机向量 X 对图 G 马尔可夫 [见式（9.3）] 时，X 的分布对图 G 因式分解 [见式（9.1）]。对该著名定理的深层次讨论见文献注释。

9.2.2 几个例子

下面提供一些例子，具体解说这些特性。

1. 离散图模型

在此先讨论离散图模型的例子，其中对应任意顶点 $s \in V$ 的随机变量 X_s 的值属于离散空间 χ_s。最简单的例子是二值情况，即 $\chi_s = \{-1, +1\}$。给定一个图 $G = (V, E)$，可以考虑概率分布族为

$$\mathbb{P}_\theta(x_1, \ldots, x_p) = \exp\left\{\sum_{s\in V} \theta_s x_s + \sum_{(s,t)\in E} \theta_{st} x_s x_t - A(\theta)\right\} \tag{9.4}$$

其参数为向量 $\theta \in \mathbb{R}^{|V|+|E|}$。为了后面推导方便，这里引入符号 $A(\theta) = \log Z(\theta)$，反映参数向量 θ 的归一化常数的依赖性。这类分布族即**伊辛模型**（Ising mode），因为它首先由伊辛（1925）用来对磁性材料的行为建模，详见文献注释。图 9-2 模拟了三种不同的伊辛模型。

伊辛模型也用于社交网络建模，例如政治家投票活动。在这种情况下，随机向量 $(X_1, \ldots, X_p)$ 表示一组 p 个政治家在一个特定法案上的投票集。假设政治家 s 在法案上投“赞成”票（$X_s = +1$）或者“反对”票（$X_s = -1$）。有 N 个法案的投票结果就可以推断 X 的联合分布。在分解式（9.4）中，参数 $\theta_s > 0$ 意味着政治家 s 在任意给定法案上更倾向（假定其他政治家的投票为固定值）于投“赞成”票，$\theta_s < 0$ 的情况下则解释相反。另一方面，对任意给定边的连接点对 s, t，其权重为 $\theta_{st} > 0$，意味着当所有政治家的投票行为固定，政治家 s 和 t 更倾向于投相同票（同时赞成或同时反对），$\theta_{st} < 0$ 的情况则相反，见后面图 9-7 的投票数据应用。

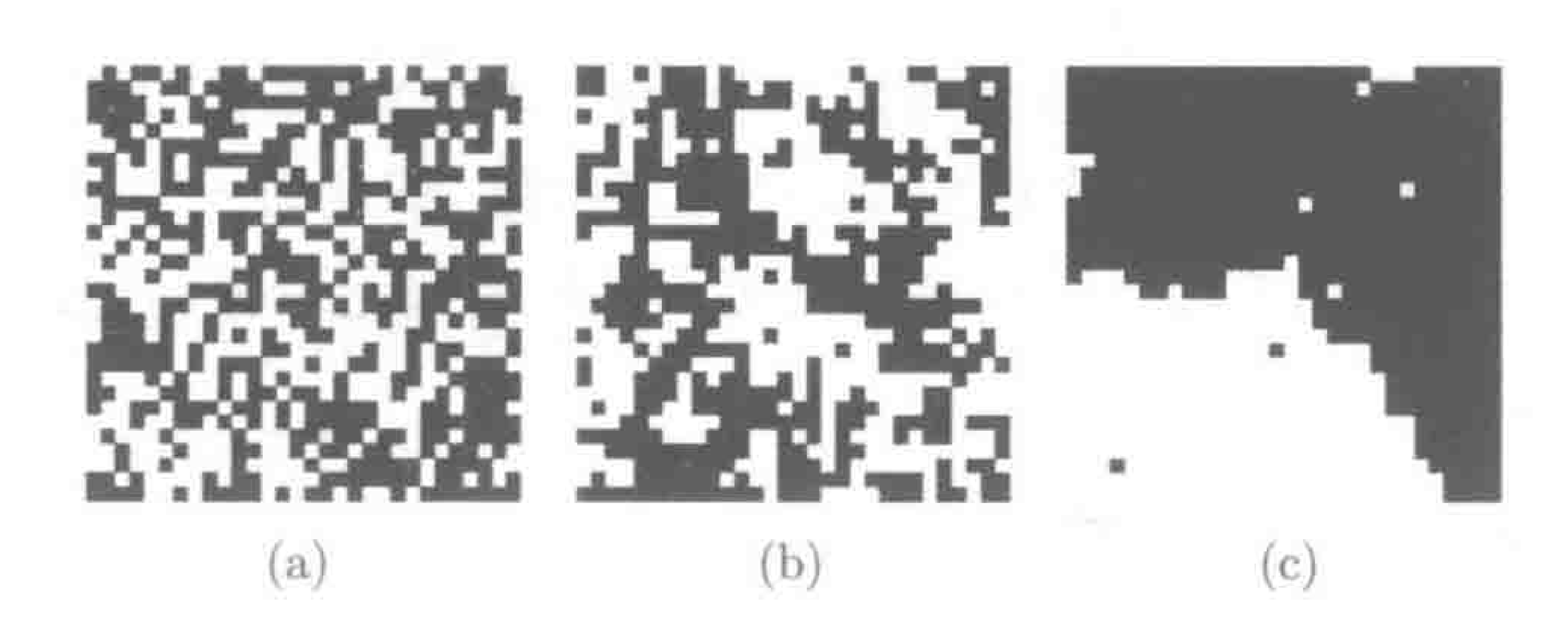


图 9-2 伊辛模型中产生的样本基于 $p=1024$ 个节点的图。为便于展示，结果向量 $x=\{-1,+1\}^{1024}$ 画在了 32×32 的二值图上。图 (a)~图 (c) 对应三种不同的分布。样本是通过 Gibbs 采样得到的

伊辛模型可以有很多扩展形式。第一，分解式 (9.4) 限制在维度最大为 2（边）的团上。允许团维度增大到 3，可以得到模型族

$$\mathbb{P}_\theta(x)=\exp\left\{\sum_{s\in V}\theta_s x_s+\sum_{(s,t)\in E}\theta_{st}x_s x_t+\sum_{(s,t,u)\in E_3}\theta_{stu}x_s x_t x_u-A(\theta)\right\} \tag{9.5}$$

其中 E_3 是一些顶点数为 3 的子集。这种分解能够扩展到高阶子集，极限情况下（同时允许所有 p 个变量之间的交互作用），能够指定任意二值向量的分布。实际上，我们感兴趣的是相对局部的交互作用模型，而非这样的全局交互模型。

伊辛模型的另一个扩展考虑了非二值变量，如 $X_s\in\{0,1,2,\ldots,m-1\}$, $m>2$。在这种情况下，关注分布族

$$\mathbb{P}_\theta(x_1,\ldots,x_p)=\exp\left\{\sum_{s\in V}\sum_{j=1}^{m-1}\theta_{s;j}\,[x_s=j]+\sum_{(s,t)\in E}\theta_{st}\,[x_s=x_t]-A(\theta)\right\} \tag{9.6}$$

其中指示函数 $\mathbb{I}[x_s=j]$ 当 $x_s=j$ 时值为 1，否则为 0。当权重 $\theta_{st}>0$ 时，基于边的指示函数 $\mathbb{I}[x_s=x_t]$ 如同一个光滑先验，相同时赋予对 (x_s,x_t) 更高权重。模型 (9.6) 在计算机视觉中很实用，比如，可应用于图像降噪和视差计算。

这里讨论的所有模型在统计学和生物统计学文献中都有，可应用于多路表上的对数线性模型。但是，在这种情况下，变量的数目非常小。

9.4.3 节将讨论混合数据下的一大类马尔可夫成对模型，同时允许离散和连续变量。

2. 高斯图模型

设 $X\sim N(\mu,\boldsymbol{\Sigma})$ 为 p 维上的高斯分布，均值向量 $\mu\in\mathbb{R}^p$，协方差矩阵 $\boldsymbol{\Sigma}$ 有

$$\mathbb{P}_{\mu,\boldsymbol{\Sigma}(x)}=\frac{1}{(2\pi)^{\frac{p}{2}}\det[\boldsymbol{\Sigma}]^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)} \tag{9.7}$$

如果将多维高斯分布视为指数族分布的一个特例，$(\mu, \boldsymbol{\Sigma})$ 则是分布族的均值参数。为了将多维高斯分布表达成一个图模型，用典型参数（即向量 $\boldsymbol{\gamma} \in \mathbb{R}^p$ 和 $\boldsymbol{\Theta} \in \mathbb{R}^{p\times p}$）进行参数化表示会更加方便。任意非退化多维高斯（即无论何时 $\boldsymbol{\Sigma}$ 都是严格正定的）可以表示为

$$\mathbb{P}_{\gamma,\boldsymbol{\Theta}(x)} = \exp\left\{\sum_{s=1}^{p}\gamma_s x_s - \frac{1}{2}\sum_{s,t=1}^{p}\theta_{st}x_s x_t - A(\boldsymbol{\Theta})\right\} \tag{9.8}$$

其中 $A(\boldsymbol{\Theta}) = -\frac{1}{2}\log\det[\boldsymbol{\Theta}/(2\pi)]$，所以 $\int \mathbb{P}_{\boldsymbol{\gamma},\boldsymbol{\Theta}}(x)dx = 1$。我们选择在分解式（9.8）中用 $-1/2$ 来归一化是为了确保矩阵 $\boldsymbol{\Theta}$ 有具体的解释。特别是，这个尺度下（见习题 9.1）有关系 $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$，所以 $\boldsymbol{\Theta}$ 对应协方差的逆，即精确度或者精度矩阵。

表达式（9.8）尤其方便，因为允许直接以精度矩阵 $\boldsymbol{\Theta}$ 的稀疏形式来讨论分解式特性。无论 X 如何根据图 G 分解，基于分解式（9.8），对任意一对 $(s,t) \notin E$ 必定有 $\boldsymbol{\Theta}_{st} = 0$，这就在 $\boldsymbol{\Theta}$ 的零模式和潜在图的边结构 E 之间建立了对应。这种对应关系的说明见图 9-3。

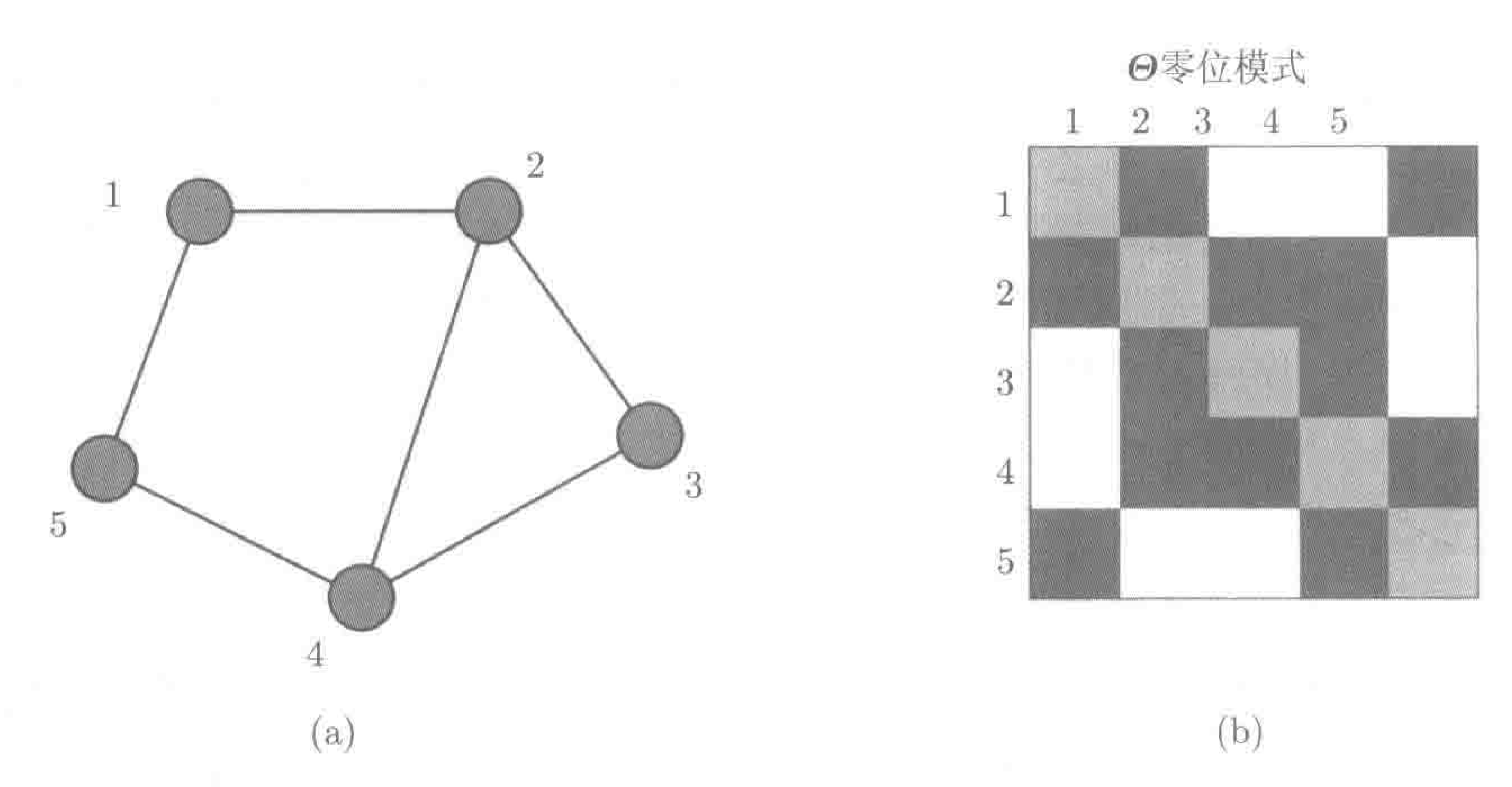


图 9-3　(a) 拥有 5 个顶点的无向图。(b) 精度矩阵 $\boldsymbol{\Theta}$ 的关联稀疏模式。白色方块对应为零项

9.3　基于惩罚似然的图选择

我们现在转向图选择问题，这个问题可以用 ℓ_1 正则化似然函数方法求解。问题简单来说就是：要从一个给定的图模型采样得到一个样本集 $\boldsymbol{X} = \{x_1, \ldots, x_N\}$，但是潜在的图的结构是未知的。怎样运用数据去选择高概率的正确图呢？这里我们讨论基于似然概率和 ℓ_1 正则化的方法。本节讨论基于图模型的全局似然函数方法。在高斯情况下，这种模型选择方法是基于带 ℓ_1 正则化的 log 行列式的凸问题，

问题易于处理。另一方面，在离散情况下，这个方法只是易于针对小图或者有特定结构的图进行计算。

9.3.1 高斯模型的全局似然性

本节讨论高斯图模型的模型选择问题，即协方差选择问题。这里的重点是估计图结构，所以假设分布有零均值。式 (9.8) 参数化后，我们只需要考虑对称精度矩阵 $\boldsymbol{\Theta} \in \mathbb{R}^{p\times p}$。

设 $\boldsymbol{X}$ 表示从零均值且精度矩阵为 $\boldsymbol{\Theta}$ 的多维高斯分布中采样得到的样本。基于一些简单的代数运算（详见习题 9.2）可以看出尺度变换后，多维高斯变量的对数似然函数 $\mathcal{L}(\boldsymbol{\Theta};\boldsymbol{X})$ 的形式为

$$L(\boldsymbol{\Theta};\boldsymbol{X}) = \frac{1}{N}\sum_{i=1}^{N}\log\mathbb{P}_{\boldsymbol{\Theta}}(x_i) = \log\det\boldsymbol{\Theta} - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Theta}) \tag{9.9}$$

其中 $\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^{\mathrm{T}}$ 为经验协方差矩阵。log det 函数在对称矩阵空间上定义为

$$\log\det(\boldsymbol{\Theta}) = \begin{cases} \displaystyle\sum_{j=1}^{p}\log(\lambda_j(\boldsymbol{\Theta})), & \boldsymbol{\Theta}\succ \boldsymbol{0} \\ -\infty, & \text{其他} \end{cases} \tag{9.10}$$

其中 $\lambda_j(\boldsymbol{\Theta})$ 是 $\boldsymbol{\Theta}$ 的第 j 个特征值。习题 9.2 探讨了这个函数的一些附加特性。目标函数 (9.9) 是 log det 规划的一个实例，log det 函数是一个研究得很透彻的优化问题。它是严格凹的，所以最大值只要存在就必定唯一，定义最大似然概率估计为 $\widehat{\boldsymbol{\Theta}}_{\mathrm{ML}}$，简记为 MLE。

根据经典理论，当样本大小 N 趋向于无穷时，最大似然估计 $\widehat{\boldsymbol{\Theta}}_{\mathrm{ML}}$ 收敛于真实精度矩阵。因此，至少理论上可以运用 $\widehat{\boldsymbol{\Theta}}_{\mathrm{ML}}$ 的阈值版本来指定一个边集，然后进行高斯图模型选择。但现实操作时常出现问题。节点 p 的个数与样本大小 N 相仿或者更大，这种情况下不存在 MLE。实际上，只要 $N<p$，经验相关矩阵 $\boldsymbol{S}$ 必定是秩退化的，这就意味着 MLE 不存在 [见习题 9.2(c)]。因此，我们必须考虑 MLE 合适的约束或者正则化形式。而且，不管样本大小如何，我们更喜欢将精度矩阵估计约束为稀疏的，这样更容易解释。

基于相对稀疏的图来求解高斯图模型，这样方便控制边的数目，可以通过基于 ℓ_0 的数值来计算

$$\rho_0(\boldsymbol{\Theta}) = \sum_{s\neq t}\mathbb{I}[\theta_{st}\neq 0] \tag{9.11}$$

其中 $\mathbb{I}[\theta_{st}\neq 0]$ 是一个 0-1 值的指示函数。注意，通过构造，我们有 $\rho_0(\boldsymbol{\Theta}) = 2|E|$，其中 $|E|$ 表示为 $\boldsymbol{\Theta}$ 图中边的数目。下面再来考虑优化问题

$$\widehat{\boldsymbol{\Theta}} \in \arg\max_{\substack{\boldsymbol{\Theta}\succeq\boldsymbol{0}\\ \rho_0(\boldsymbol{\Theta})\leqslant k}}\{\log\det(\boldsymbol{\Theta}) - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Theta})\} \tag{9.12}$$

遗憾的是，基于 ℓ_0 的约束定义了一个高度非凸约束集，集合为所有包含 k 条边的 $\left(\binom{\binom{p}{2}}{k}\right)$ 个可能子集。

因此，我们很自然地考虑凸松弛，用对应的 ℓ_1 约束替换 ℓ_0 约束。这样得到凸规划

$$\hat{\boldsymbol{\Theta}} \in \arg\max_{\boldsymbol{\Theta}\succeq 0} \{\log\det(\boldsymbol{\Theta}) - \text{trace}(\boldsymbol{S\Theta}) - \lambda\rho_1(\boldsymbol{\Theta})\} \tag{9.13}$$

其中 $\rho_1(\boldsymbol{\Theta}) = \sum_{s\neq t}|\theta_{st}|$ 是 $\boldsymbol{\Theta}$ 的非对角线元素的 ℓ_1 形式。问题（9.13）可以视为 log-det 规划的一个实例。因此这是一个凸规划，常记为图 lasso。

因为这是一个凸规划，所以可以用常见的内点法来求解，参见 Vandenberghe, Boyd and Wu(1998)。但是，这对大型问题效率并不高。更自然的方法是一阶块坐标下降法，由 d'Aspremont, Banerjee and El Ghaoui（2008）引入，Friedman, Hastie and Tibshirani （2008）精简。后者称为**图 lasso算法**。此算法有一个简单的形式，与基于近邻的回归方法有关联，这将在 9.4 节讨论。

9.3.2 图 lasso 算法

式（9.13）的次梯度公式为

$$\boldsymbol{\Theta}^{-1} - \boldsymbol{S} - \lambda\cdot\boldsymbol{\Psi} = \boldsymbol{0} \tag{9.14}$$

其中对称矩阵 $\boldsymbol{\Psi}$ 的对角线元素均为 0。如果 $\theta_{jk}\neq 0$，则 $\psi_{jk} = \text{sgn}(\theta_{jk})$；如果 θ_{jk}=0，则 $\psi_{jk}\in[-1,1]$。

我们现在考虑通过块坐标下降法来求解该问题。为此，我们考虑分割所有的矩阵为一列和剩余部分，为了方便，这里取最后一列：

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}^T & \boldsymbol{\theta}_{22} \end{bmatrix},\ \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^T & \boldsymbol{s}_{22} \end{bmatrix},\text{等等} \tag{9.15}$$

$\boldsymbol{\Theta}^{-1}$ 的当前版本记为 $\boldsymbol{W}$，分割为（9.15），然后固定除最后一行和一列以外的剩余部分。运用分区后的逆，式（9.14）变为

$$\boldsymbol{W}_{11}\beta - \boldsymbol{s}_{12} + \lambda\cdot\boldsymbol{\psi}_{12} = \boldsymbol{0} \tag{9.16}$$

其中 $\beta = -\boldsymbol{\theta}_{12}/\boldsymbol{\theta}_{22}$。这里已经固定了第 p 行和第 p 列：$\boldsymbol{W}_{11}$ 是 $\boldsymbol{\Theta}^{-1}$ 的 $(p-1)\times(p-1)$ 块，$\boldsymbol{s}_{12}$ 和 $\boldsymbol{\theta}_{12}$ 是 $\boldsymbol{S}$ 和 $\boldsymbol{\Theta}$ 的第 p 行和第 p 列的 $p-1$ 个非对角线元素。$\boldsymbol{\theta}_{22}$ 是 $\boldsymbol{\Theta}$ 的第 p 个对角线元素。这些细节在习题 9.6 中推导。①

① 以往有记录显示，这一算法并非针对式（9.14）的 $\boldsymbol{\Theta}$ 的块坐标下降，而是相当于解决式（9.13）中凸对偶问题的块坐标下降步骤。这在 Banerjee, El Ghaoui and d'Aspremont （2008） 中有暗示，Mazumder and Hastie(2012) 中有详细说明。对偶变量为 $\boldsymbol{W} = \boldsymbol{\Theta}^{-1}$。后者对原问题推导出了另一种坐标下降算法（见习题 9.7）。在一些情况下，这比原始图 lasso 算法数值表现更好。

可以看出，式 (9.16) 等价于 lasso 回归估计方程的改进版。考虑到通常的回归有输出变量 $\boldsymbol{y}$ 和预测子矩阵 $\boldsymbol{Z}$。在这个问题中，lasso 最小化

$$\frac{1}{2N}(\boldsymbol{y}-\boldsymbol{Z}\beta)^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{Z}\beta)+\lambda\cdot\|\beta\|_1 \tag{9.17}$$

则次梯度公式为

$$\frac{1}{N}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\beta-\frac{1}{N}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}+\lambda\cdot\mathrm{sgn}(\beta)=0 \tag{9.18}$$

相比式 (9.16)，可见 $\frac{1}{N}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}$ 对应于 $\boldsymbol{s}_{12}$，$\frac{1}{N}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ 对应于 $\boldsymbol{W}_{11}$。$\boldsymbol{W}_{11}$ 是从当前模型中得到的向量积矩阵估计。因此，我们可以用改进 lasso 算法逐块求解式 (9.16)，即将各变量视为响应变量，其余的 $p-1$ 个变量视为预测子。总结见算法 9.1。

Friedman et al. (2008) 在各步中运用逐路径坐标下降方法求解改进版 lasso 问题（对 λ 降序）。[这对应于 lasso 算法的 "协方差" 版本，与 R 和 matlab 里 glmnet 包中用的一样 (Friedmanet al. 2010b)。]

从式 (9.14) 中可以看到, 解矩阵 $\boldsymbol{W}$ 的对角线元素 w_{jj} 是 s_{jj}，在算法 9.1 的第 1 步中已固定。[①]

算法 9.1 图 lasso

---

1. 初始化 $\boldsymbol{W}=\boldsymbol{S}$。注意，$\boldsymbol{W}$ 的对角线元素在下面步骤中不变。
2. 对 j=1,2,. . . p,1,2,. . . p, . . . 重复直到收敛
   (a) 分割矩阵 $\boldsymbol{W}$ 为两部分，一部分是除 j 行和列外的所有，另一部分是第 j 行和列。
   (b) 用求解改进版 lasso 的循环坐标下降算法求解估计公式 $\boldsymbol{W}_{11}\beta-\boldsymbol{s}_{12}+\lambda\cdot\mathrm{sgn}\,(\beta)=0$。
   (c) 迭代 $\boldsymbol{w}_{12}=\boldsymbol{W}_{11}\widehat{\boldsymbol{\beta}}$。
3. 在最后一次循环（对各个 j）求解 $\widehat{\boldsymbol{\theta}}_{12}=-\widehat{\beta}\cdot\widehat{\boldsymbol{\theta}}_{22}$，其中 $1/\widehat{\boldsymbol{\theta}}_{22}\boldsymbol{w}_{22}-\boldsymbol{w}_{12}^{\mathrm{T}}\widehat{\beta}$。

图 lasso 算法速度很快，可以在 1 分钟内求解一个带 1000 个节点的中等稀疏问题。这很容易修改算法，使具体的边有惩罚系数 λ_{jk}。注意，$\lambda_{jk}=\infty$ 将会迫使 $\widehat{\theta}_{jk}$ 为 0。

图 9-4 用一个简单的例子说明了路径算法。图 9-3 的模型可以生成 20 组样本，有

$$\boldsymbol{\Theta}=\begin{bmatrix}2&0.6&0&0&0.5\\0.6&2&-0.4&0.3&0\\0&-0.4&2&-0.2&0\\0&0.3&-0.2&2&0\\0.5&0&0&-0.2&2\end{bmatrix} \tag{9.19}$$

---

① 这里提出该问题的另一种提法，惩罚 $\boldsymbol{\Theta}$ 的对角项和非对角项。这样一来，解矩阵的对角项 w_{jj} 为 $s_{jj}+\lambda$，算法的其余部分不变。

图为 λ 在一定范围内的图 lasso 估计，横轴为解 $\widehat{\boldsymbol{\Theta}}_\lambda$ 的 ℓ_1 范数。$\boldsymbol{\Theta}$ 的真实值在图的右方。解集很好，但是因为右边的解是 $\boldsymbol{S}^{-1}$（没有正则化），等于 $\boldsymbol{\Theta}$（见习题 9.3），所以这并不奇怪。在右图中，数据矩阵的各项加上了一个独立高斯噪声（标准差 0.05）。现在可以看到，估计几乎不精确了。事实上，非零支撑项在解径上的值从未准确过。

下面列举一些关于高斯图模型的深层观点。

- 有一个简单的方法是用观测方差 $\boldsymbol{S}_{11}$ 代替 $\boldsymbol{W}_{11}$。这只需要一次遍历预测子，将各个变量 X_j 对其他变量做 lasso 回归。这称为**近邻选择**（Meinshausen and Bühlmann 2006）。同图 lasso 算法一样，这样得到的是 $\boldsymbol{\Theta}$ 支撑的一致估计，但并不保证得到的是正定估计 $\widehat{\boldsymbol{\Theta}}$。我们将在 9.4 节中详细讨论这一点。
- 如果在 $\boldsymbol{\Theta}$ 中预先指定零模式，则可以用标准线性回归替代 lasso，不计算预计系数为零的预测子。这将提供一个方便的方法计算带约束的 $\boldsymbol{\Theta}$ 的最大似然估计。详见 Hastie et al.（2009）Chapter 17。

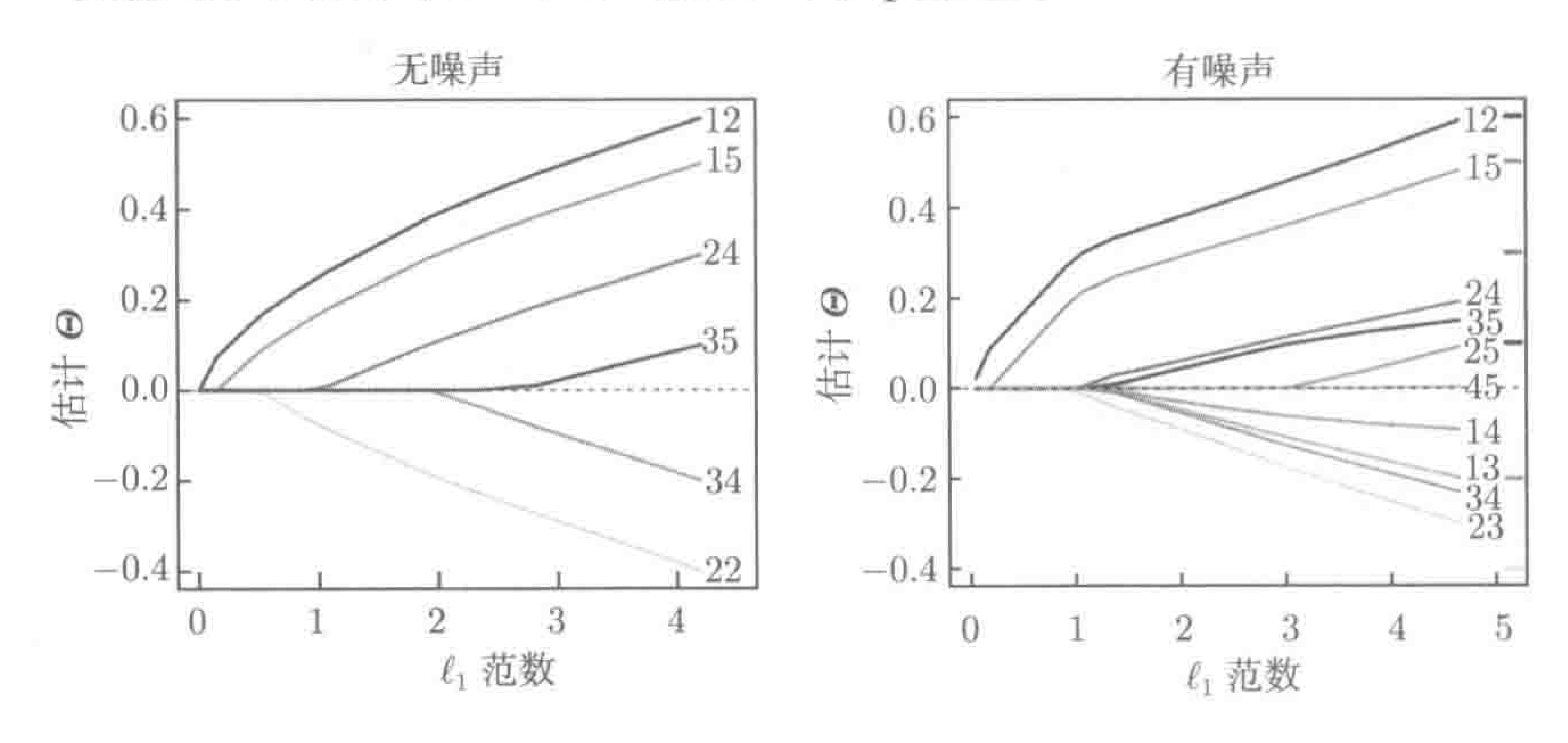


图 9-4　左图：图 lasso 的估计，数据为图 9-3 中的模型模拟得到。$\boldsymbol{\Theta}$ 的真实值在图的最右边。右图：相同的步骤，只是在数据的每一列中加入了标准高斯噪声。沿着路径没有一处是真正的边缘集

9.3.3 利用块对角化结构

如果逆协方差矩阵对一些有序变量有块对角化结构

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Theta}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Theta}_{kk} \end{bmatrix} \tag{9.20}$$

则图 lasso 问题可以按照各块单独求解，解可以从各个单独解上构建。这一点可以直接从次梯度式（9.14）得到。

事实证明，有一个十分简单的充分必要条件可以使图 lasso 的解拥有这种结构（Witten, Friedman and Simon 2011，Mazumder and Hastie 2012）。设 $C_1, C_2, \ldots, C_K$ 把编号为 $1,2,\ldots,p$ 的 $\boldsymbol{S}$ 分解成的 K 块。则当且仅当对不属于相同块的所有对 (i,i') 有 $|s_{ii'}| \leqslant \lambda$ 时，$\widehat{\boldsymbol{\Theta}}$ 的对应部分有相同的块结构。证明很简单，通过式（9.14），运用定理“块对角矩阵的逆矩阵有相同的块对角结构”即可。这意味着，各个块 C_K 的元素与其他块的元素之间毫无关系。

这个理论可以用来使计算显著加速，做法是首先确定无联系的元素，然后求解各块中的子问题。块的数目随 λ 单调。这意味着，只要 λ 足够大，一般难以求解的大问题都可以得到解。

9.3.4 图 lasso 的理论保证

为了探讨图 lasso 特性即 log-det 方法，我们进行了一系列的仿真。对于一个有 p 个节点和指定协方差矩阵的图，产生了一组 N 个零均值多维高斯分布样本，然后再用样本产生经验协方差矩阵 $\boldsymbol{S}$。在此用正则化参数 $\lambda_N = 2\sqrt{\frac{\log p}{N}}$（Ravikumar, Wainwright, Raskutti and Yu 2011）求解图 lasso 问题，然后画出算子范式误差 $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2$，横轴是样本大小为 N 的图。图 9-5（左）所示为二维网格图的三种不同维度 $p \in \{64, 100, 225\}$的情况，其中各个节点的自由度为 4。对于不同的图维度，我们得到的协方差矩阵的逆矩阵 $\boldsymbol{\Theta}^* \in \mathbb{R}^{p\times p}$，各项为

$$\theta^*_{st} = \begin{cases} 1, & s=t \\ 0.2, & |s-t|=1 \\ 0, & \text{其他} \end{cases}$$

图中的左边部分证明，图 lasso 对于算子范式中的协方差逆 $\boldsymbol{\Theta}^*$ 的估计是一个一致过程，因为其误差曲线随着 N 的增加收敛于零。比较不同大小图的曲线，我们可以看到误差曲线向右上方移动，反映了大的图需要更多的样本来达到相同的容错率。

众所周知，图 lasso 的解 $\widehat{\boldsymbol{\Theta}}$（以高概率）满足误差范围

$$\left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_2 \lesssim \sqrt{\frac{d^2 \log p}{N}} \tag{9.21}$$

其中 d 是图中任意节点的最大自由度，$\lesssim$ 表示不等式最大到常数项（详见参考文献注释）。如果理论预测准确，则我们可能希望相同的误差曲线相对符合。（若重新画图，采用尺度调整后的样本大小（$\frac{N}{d^2 \log p}$。）图 9-5 的右半部分为这一预测的经验验

证。注意，也有理论结果保证，只要 $N = \Omega\left(d^2 \log p\right)$，图 lasso 估计的支撑集 $\widehat{\Theta}$ 就与 Θ^* 的支撑集相符合。因此，图 lasso 也能成功得到真实的图结构。

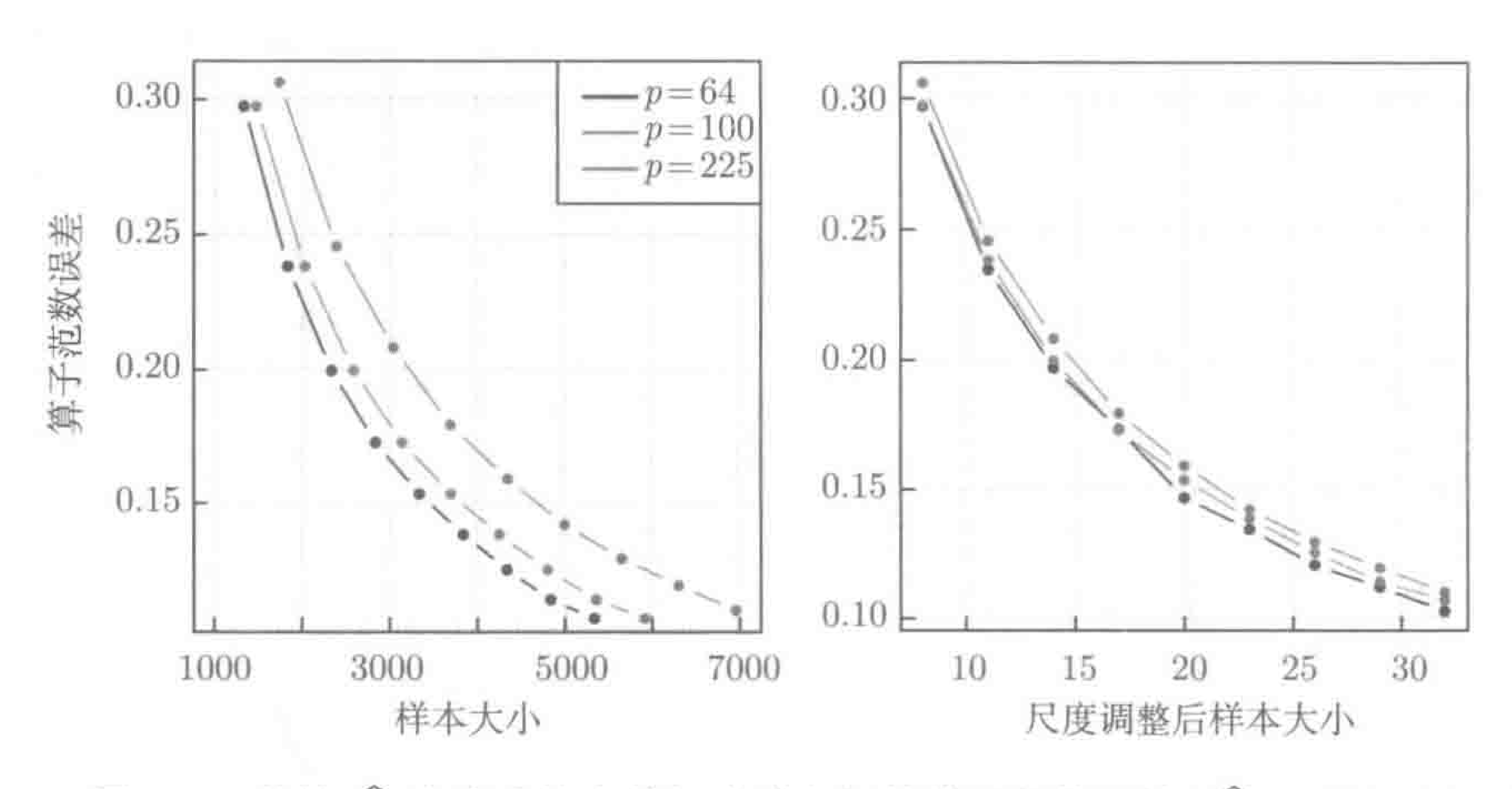


图 9-5　图 lasso 估计 $\widehat{\Theta}$ 和真实协方差矩阵逆之间的算子范数误差 $\|\widehat{\Theta}-\Theta^*\|_2$ 图。左图：横轴为原始样本大小 N，三条线 $p \in\{64, 100, 225\}$。注意曲线如何随着图维度 p 增加转移到右边，这体现了一个定理：大图的一致估计需要更多的样本。右图：同样为算子–范数误差曲线，横轴为尺度调整后样本大小 $\frac{N}{d^2 \log p}$，三条线 $p \in\{64, 100, 225\}$。根据相关理论可知，曲线对齐得很好

9.3.5　离散模型的全局似然性

理论上讲，在带离散变量的图模型选择问题上，可以设想采用带 ℓ_1 正则化的全局似然概率。这里的一个主要挑战式（9.4）至式（9.6）中的配分函数 A 通常很难计算，这与多维高斯分布形成了鲜明对比。例如，在伊辛模型（9.4）中，其形式为

$$A(\theta) = \log\left[\sum_{x \in\{-1,+1\}^p} \exp\left\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\right\}\right]$$

因此，暴力方法对大的 p 比较棘手，因为其中有一个包括 2^p 项的求和。除了一些特殊情况，$A(\theta)$ 的值整体上难以计算。

有多种方法可以近似配分函数（见参考文献）。但是，这些方法多少远离了本章讨论的主要内容。下一节会从全局似然概率转向讨论条件或者伪似然概率。这些方法可以同时运用于高斯和离散变量模型，从而方便计算。这意味着存在多项式时间，无关样本大小和图的大小。

9.4 基于条件推断的图选择

图选择的一个替代算法是基于近邻似然概率的思路，或者伪似然概率的乘积。这些算法都关注条件分布，条件分布在很多情况下易于处理。

对于一个给定的顶点 $s \in V$，图中其余随机变量的集合可以表示为

$$X_{\backslash\{s\}} = \{X_t, t \in V\backslash\{s\}\} \in \mathbb{R}^{p-1}$$

现在考虑给定随机向量 $X_{\backslash\{s\}}$ 下 X_s 的分布。通过任意无向图模型（见 9.2 节）的条件独立特性可知，这一条件的唯一相关变量在相邻集中

$$\mathcal{N}(s) = \{t \in V \mid (s,t) \in E\} \tag{9.22}$$

事实上，如图 9-6 所示，集合 $\mathcal{N}(s)$ 为一割集，分割了 $\{s\}$ 和剩余顶点 $V\backslash\mathcal{N}^+(s)$，在此定义 $\mathcal{N}^+(s) = \mathcal{N}(s) \cup \{s\}$。因此，可以保证变量 X_s 在给定变量 $X_{\mathcal{N}(s)}$ 下条件独立于 $X_{V\backslash\mathcal{N}^+(s)}$，或者等价于

$$(X_s \mid X_{\backslash\{s\}}) \stackrel{d}{=} (X_s \mid X_{\mathcal{N}(s)}) \tag{9.23}$$

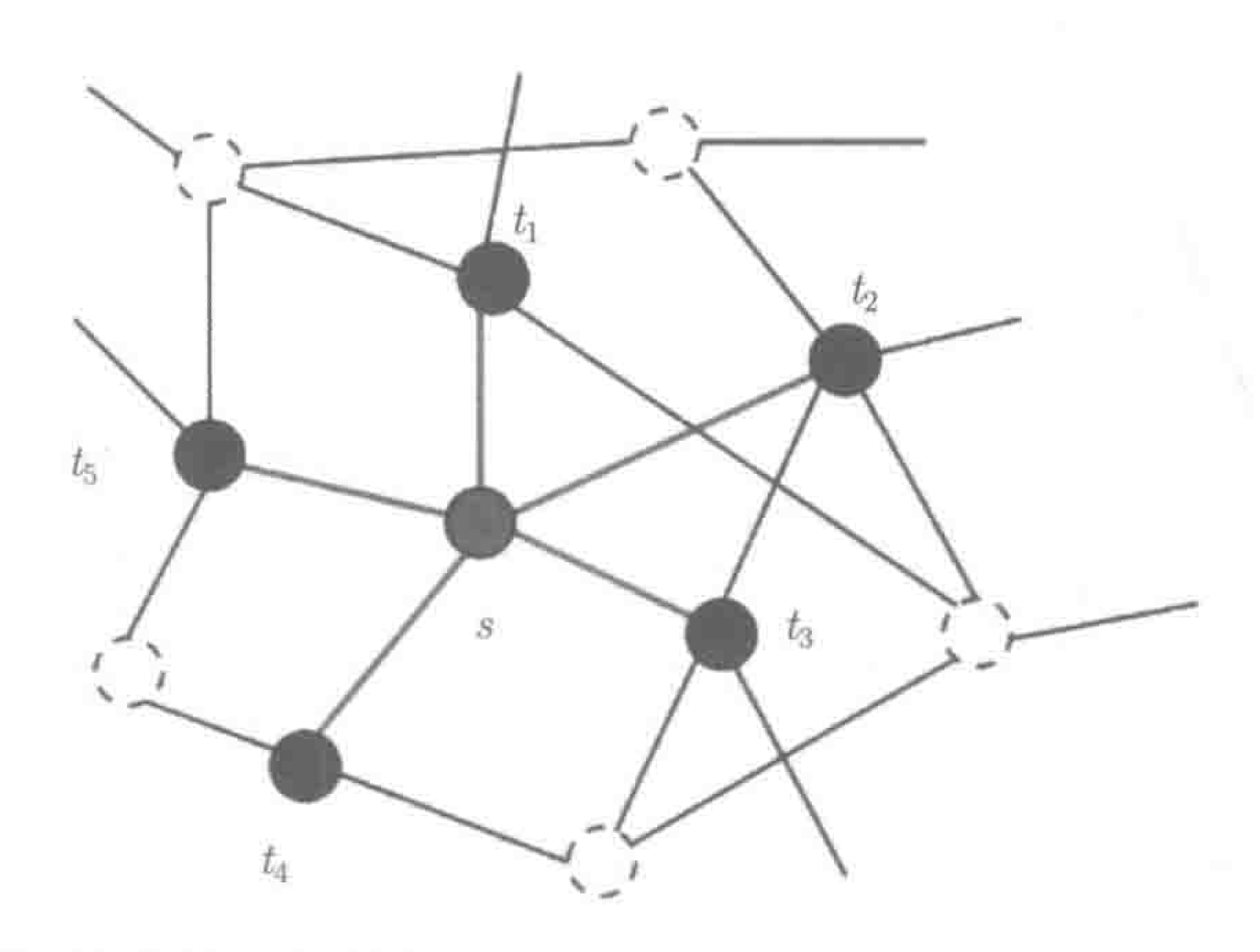


图 9-6 深蓝色顶点形成了红色顶点的相邻集 $\mathcal{N}(s)$；集合 $\mathcal{N}^+(s)$ 由 $\mathcal{N}(s) \cup \{s\}$ 联合而成。注意，$\mathcal{N}(s)$ 是图中的割集，分割 $\{s\}$ 和 $V\backslash\mathcal{N}^+(s)$。因此，变量 X_s 在给定相邻集中变量 $X_{\mathcal{N}(s)}$ 下条件独立于 $X_{V\backslash\mathcal{N}^+(s)}$。这种条件独立性意味着基于图中其他所有变量的 X_s 的最优预测只依赖于 $X_{\mathcal{N}(s)}$（见彩插）

怎样在图选择上利用这种条件独立（Conditional Independence，CI）特性呢？如果考虑基于 $X_{V\backslash\{s\}}$ 预测 X_s 的值问题，那么由 CI 特性可知，最好的预测子可以通过只含 $X_{\mathcal{N}(s)}$ 的函数指定。因此，寻找相邻集问题可以通过求解预测问题来解决。

9.4.1 高斯分布下基于近邻的似然概率

本节首先介绍多维高斯分布下的这种方法。在这种情况下，给定 $X_{\backslash\{s\}}$，X_s 的条件分布依然是高斯分布，所有 X_s 都可以分解为最佳线性预测，依据是 $\boldsymbol{X}_{\backslash\{s\}}$ 和一个误差项，即

$$X_s = X_{\backslash\{s\}}^{\mathrm{T}}\beta^s + W_{\backslash\{s\}} \tag{9.24}$$

在这种分解下，$W_{\backslash\{s\}}$ 是一个零均值高斯变量，$\mathrm{Var}(W_{\backslash\{s\}})=\mathrm{Var}(X_s|X_{\backslash\{s\}})$，对应于预测误差，独立于 $X_{\backslash\{s\}}$。因此，这种依赖性完全为线性回归系数 β^s 所描述，β^s 由 θ^s 乘以一个标量得到，对应于 9.2.2 节式（9.8）中 $\boldsymbol{\Theta}$ 的子向量（见习题 9.4）。

对式（9.24）进行分解，可以证明在多维高斯分布下，预测问题降为 X_s 在 $X_{\backslash\{s\}}$ 上的线性回归问题。这里最主要的特性（如习题 9.4 所示），是回归向量 β^s 满足 $\sup\mathrm{p}(\beta^s)=\mathcal{N}(s)$。如果图相对稀疏，即节点 s 的自由度 $|\mathcal{N}(s)|$ 相对于 p 较小，那么这里自然要考虑通过 lasso 估计 β^s。这就引出了高斯图模型选择中下面这种基于近邻的方法，其样本为 $\boldsymbol{X}=\{x_1,\ldots,x_N\}$。

在步骤 1a 中，$x_{i,V\backslash\{s\}}$ 表示 p 维向量 x_i 的 $(p-1)$ 维子向量，忽略了第 s 个元素。为了说清步骤 2，AND 规则为当且仅当 $s\in\widehat{\mathcal{N}}(t)$ 和 $t\in\widehat{\mathcal{N}}(s)$，边 (s,t) 属于边估计集 $\widehat{E}$。另一方面，OR 规则不太严格，任意 $s\in\widehat{\mathcal{N}}(t)$ 或者 $t\in\widehat{\mathcal{N}}(s)$ 下都有 $(s,t)\in\widehat{E}$。

算法 9.2 基于近邻的高斯图模型选择

---

1. 对各个顶点 $s=1,2,\ldots,p$, 进行如下计算：
   (a) 将 lasso 用于求解近邻预测问题：
   $$\widehat{\beta}^s \in \underset{\beta^s\in\mathbb{R}^{p-1}}{\operatorname{argmin}}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(x_{is}-x_{i,V\backslash\{s\}}^{\mathrm{T}}\beta^s\right)^2+\lambda\|\beta^s\|_1\right\} \tag{9.25}$$
   (b) 计算相邻集 $\mathcal{N}(s)$ 的估计 $\widehat{\mathcal{N}}(s)=\sup\mathrm{p}(\widehat{\beta}^s)$。
2. 通过 AND 或者 OR 规则综合相邻集估计 $\{\widehat{\mathcal{N}}(s),s\in V\}$，形成图的估计 $\widehat{G}=(V,\widehat{E})$。

相邻集模型的优点是速度快。lasso 有很多高效实现，p 回归问题可以独立求解，因此可以并行计算。

AND/OR 规则可以用基于伪似然概率的联合估计方法避免，伪似然概率本质上是式（9.25）中 log 似然概率的和。本例利用 $\boldsymbol{\Theta}$ 的对称性，虽然更加简洁，但是将附加一点计算成本（Friedman, Hastieand Tibshirani 2010a）。这样能得到估计 $\widehat{\boldsymbol{\Theta}}$，而不只是图结构。9.4.3 节将更深层地讨论这种方法。

9.4.2 离散模型下基于近邻的似然概率

基于近邻的似然概率的观点并不限于高斯模型，也可以运用在其他可以写成指数族形式的图模型中。事实上，给定全局似然概率，针对离散图模型的计算是非

常棘手的。这样基于近邻的方法在离散情况下则极具吸引力，至少从计算角度上讲是这样的。

最简单的离散图模型是伊辛模型 (9.4)，用于对两两相互作用的一组变量 $(X_1, X_2, \ldots, X_p) \in \{-1,+1\}^p$ 进行建模。这种情况下（见习题 9.5），给定 $X_{V\backslash\{s\}}$，X_s 概率的条件 log-odds 形如 ①

$$\eta\theta^s(X_{\backslash\{s\}}) = \log\left[\frac{\mathbb{P}(X_s=+1|X_{\backslash\{s\}})}{\mathbb{P}(X_s=-1|X_{\backslash\{s\}})}\right] = 2\theta_s + \sum_{t\in V\backslash\{s\}} 2\theta_{st}X_t \tag{9.26}$$

其中 $\theta^s = \left[\theta_s, \{\theta_{st}\}_{t\in V\backslash\{s\}}\right]$。因此，伊辛模型的基于近邻的方法有与算法 9.2 相同的形式，其中步骤 1a 中的普通 lasso 替换为 lasso 逻辑斯蒂回归问题

$$\hat{\theta}^s \in \underset{\theta^s\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\ell\left[x_{is}, \eta\theta^s(\boldsymbol{x}_{i,\backslash\{s\}})\right] + \lambda\sum_{t\in V\backslash\{s\}}|\theta_{st}|\right\} \tag{9.27}$$

其中 ℓ 是二项式分布下的负对数似然概率。这又是一个凸问题，任意为 ℓ_1 惩罚逻辑斯蒂回归设计的算法都可以用，比如第 5 章中讨论的坐标下降算法。在高斯情况下，规则必须保证边的对称性。

Hoefling and Tibshirani（2009）针对这一问题提出了一种伪似然概率方法，保证了对称性。这可以视作上面描述的全局似然概率方法和基于近邻方法的中间产物，但是计算起来很棘手。9.4.3 节会介绍这一方法。

对于政客投票数据重建，用伊辛模型拟合美国参议院政客社交网络可以得到图 9-7，详细描述见图标题。总体上，我们可以看到各个党派内的强关联。

可以证明，在样本数目相对合适的情况下，基于近邻的方法具有一致性（详见文献注释）。在伊辛模型下，$\widehat{G}$ 表示采用逻辑斯蒂回归的基于近邻方法的输出。众所周知，$N = \Omega\left(d^2\log p\right)$ 个样本足够以高概率保证 $\widehat{G} = G$。图 9-8 表明，当该方法用于一个带 p 个节点的星形图时，其中中心节点连接 $d = \lceil 0.1p\rceil$ 个分支节点，这个条件实际已足够。（图中剩余的节点与星形节点子图不连接。）对有 $p \in\{64, 100, 225\}$个节点的图应用基于近邻的逻辑斯蒂回归方法，用 AND 法则结合近邻，形成图估计。图 9-8a 绘出了正确得到未知图结构的概率 $\mathbb{P}[\widehat{G} = G]$，横轴为样本数目 N，各条曲线对应不同的图大小。由此可见，该方法具有模型–选择一致性，因为随着样本数目 N 的增加，概率 $\mathbb{P}[\widehat{G} = G]$ 收敛于 1。对于大的图，由失败到成功自然会更晚出现（在更大的样本数目下），这说明问题更难。图 9-8b 显示了相同的模拟结果，横轴为调

① 乘子 2 源于响应变量编码为 $+1/-1$，而非传统的 0/1。

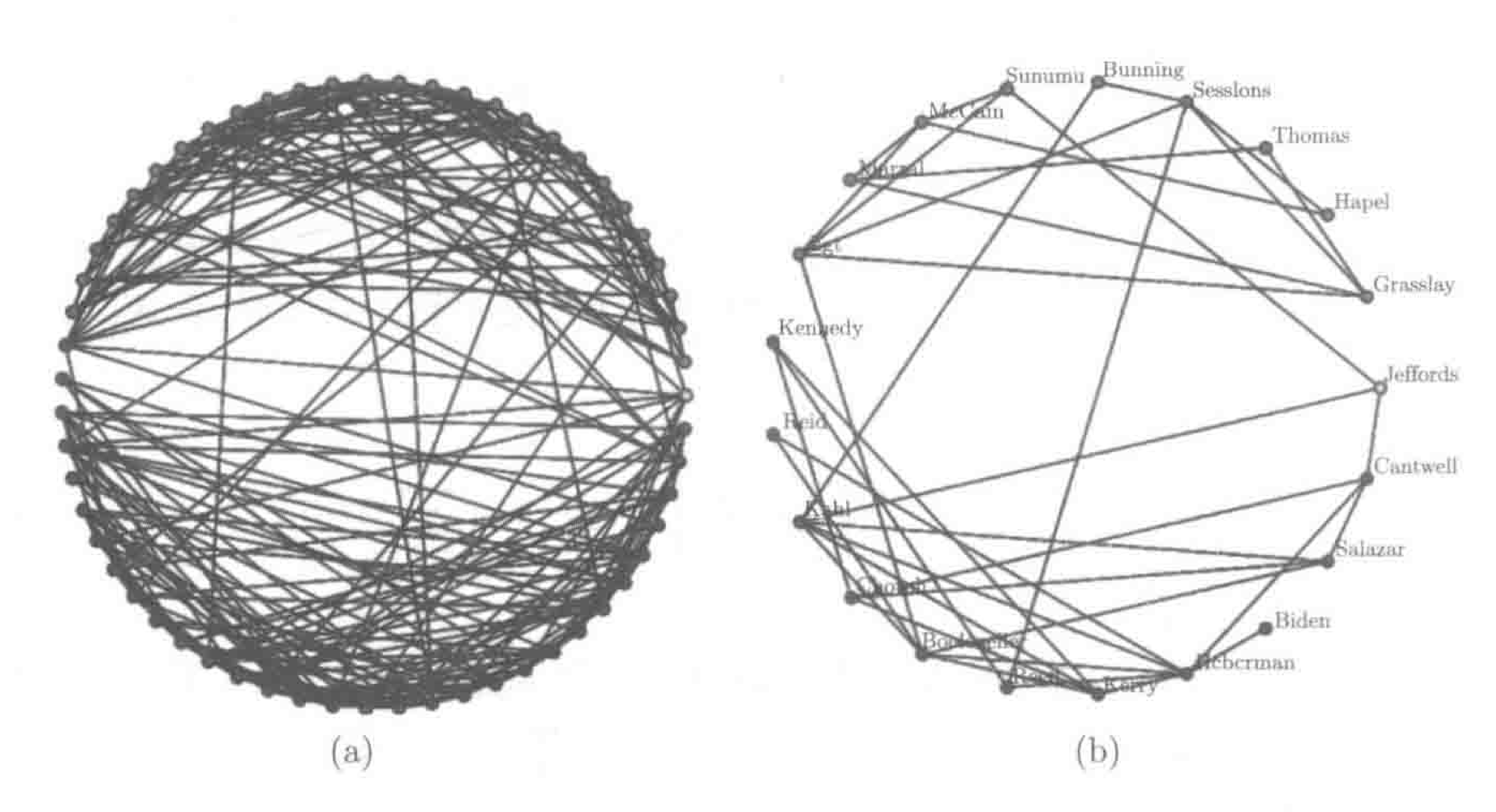


图 9-7　美国参议院（2004—2006）投票数据估计出来的政客网络。数据集为 p=100 个参议员，总共 N=546 场投票，$X_s=+1$（$X_s=-1$）意味着参议员 s 投了“赞成”（“反对”）。这里用基于近邻的逻辑斯蒂回归方法拟合数据得到一对图模型。(a) 拟合包含 55 个参议员的子图，蓝色/红色/黄色分别表示民主/共和/独立党派参议员。注意，子图显示集群根据党派有一个鲜明的两部分趋势。少量的参议员有跨党派的联系。(b) 相同社交网络的更小子图（见彩插）

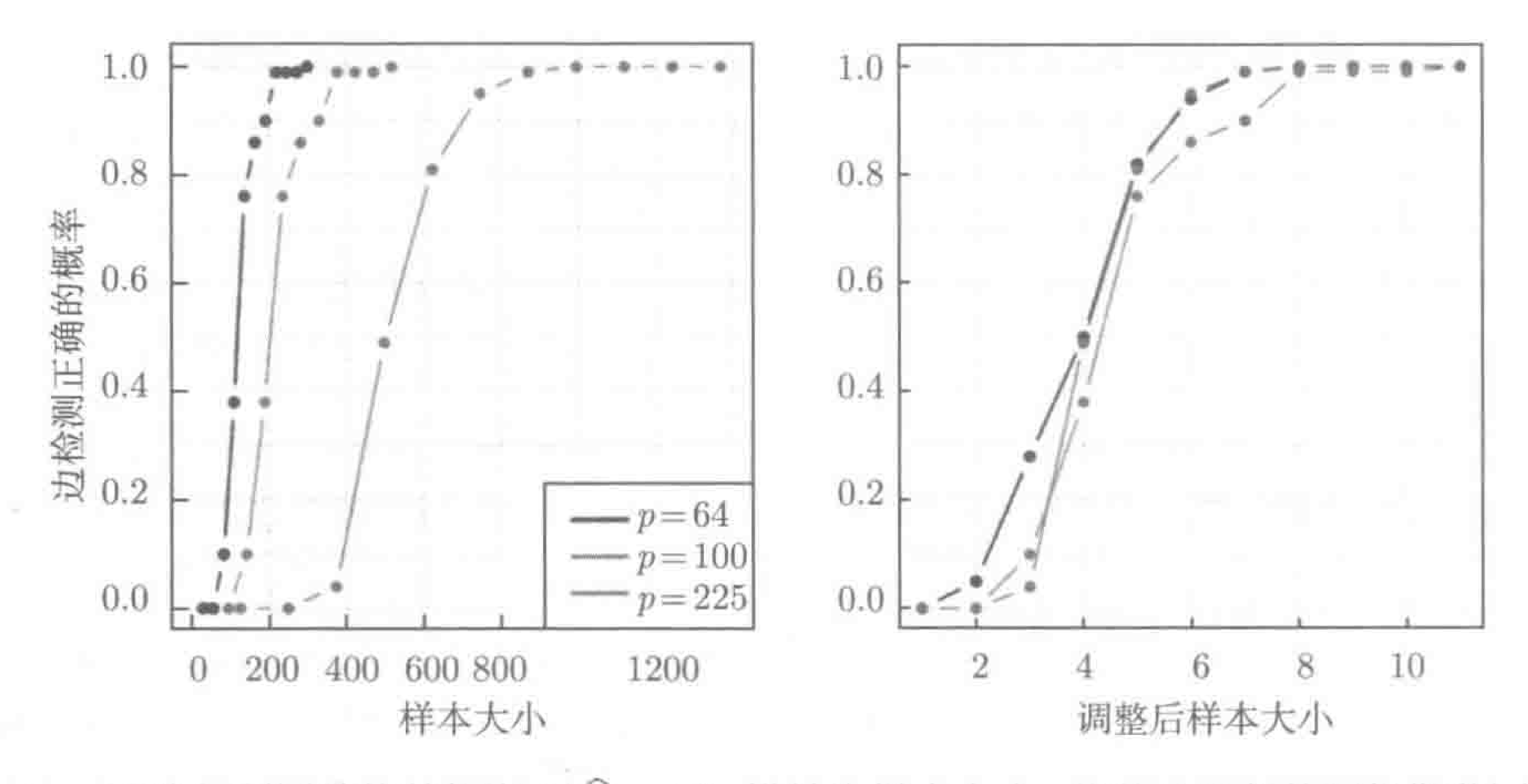


图 9-8　正确得到图结构的概率 $\mathbb{P}[\widehat{G}=G]$，横轴为样本大小。将基于近邻的逻辑斯蒂回归方法应用于估计星形图（中心辐射状），p 个顶点，中心自由度为 $d=\lceil 0.1p\rceil$。(a) 横轴为原样本数目 N。大的图一致估计需要更多的样本。(b) 相同的模拟结果，横轴为调整后样本数目 $\frac{N}{d\log p}$

整后的样本数目 $\frac{N}{d\log p}$，在新坐标系下，所有三条曲线都对齐工整。这一模拟证实了理论上的尺度 $N=\Omega\left(d^2\log p\right)$ 可以充分确保成功得到图结构，但是对这类图却不必要。然而，对其他类型的图，这个尺度是充分必要条件（详见文献注释）。

9.4.3 混合模型下的伪似然概率

目前为止，我们已经讨论了所有连续变量（高斯图模型）和所有二值变量（伊辛模型）下的图模型。没有包含其余常见情况。

(1) 离散变量（状态多于两个）

(2) 混合数据类型（例如，有连续的也有离散）

本节将之前讨论的模型扩展到以上这些例子中，演示一种基于伪似然概率的推断方法。

高斯和伊辛模型的一个简单推广就是成对马尔可夫随机场模型。为了便于表示，这里用 X 表示 p 个连续变量，Y 表示 q 个离散变量。密度 $\mathbb{P}_\Omega(x,y)$ 正比于

$$\exp\left\{\sum_{s=1}^{p}\gamma_s x_s-\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}\theta_{st}x_s x_t+\sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}[y_j]x_s+\sum_{j=1}^{q}\sum_{r=1}^{q}\psi_{jr}[y_j,y_r]\right\} \tag{9.28}$$

前两项与高斯图模型（9.8）情形相同。项 ρ_{sj} 表示连续变量 X_s 和离散变量 Y_j 之间的边。如果 Y_j 有 L_j 个可能状态或者水平，则 ρ_{sj} 为有 L_j 个参数的向量，$\rho_{sj}[y_j]$ 表示第 y_j 个值。同样，ψ_{jr} 为 $L_j\times L_r$ 矩阵，表示离散变量 Y_j 和 Y_r 之间的边，$\psi_{jr}[y_j,y_r]$ 表示 y_j 行和 y_r 列上的元素。项 ψ_{jj} 为对角元素，表示节点潜能 [对应伊辛模型（9.4）中的 θ_s]。矩阵 $\boldsymbol{\Omega}$ 表示整个参数集。不用说，配分函数一般难以处理，除非维度非常低。

伪似然概率在这里十分实用。它是 $p+q$ 条件似然概率的积，其中每一项都简单（见习题 9.8），依赖于响应项的形式。

连续：各个 p 连续变量的条件分布是高斯分布，条件变量均值具有线性。

$$\mathbb{P}(X_s|X_{\setminus\{s\}},Y;\boldsymbol{\Omega})=\left(\frac{\theta_{ss}}{2\pi}\right)^{\frac{1}{2}}\mathrm{e}^{-\frac{\theta_{ss}}{2}\left(X_s-\frac{\gamma_s+\sum_j\rho_{sj}[Y_j]-\sum_{t\neq s}\theta_{st}X_t}{\theta_{ss}}\right)^2} \tag{9.29}$$

右边离散条件变量的作用体现在不同的加法常量上，如同线性回归模型中的定性因子，例如，各个水平中的常量由 ρ_{sj} 决定。

离散：各个 q 离散变量的条件分布是多项式分布，条件变量具有 log 概率线性。

$$\mathbb{P}(Y_j|X,Y_{\setminus\{j\}};\boldsymbol{\Omega})=\frac{\mathrm{e}^{\psi_{jj}[Y_j,Y_j]+\sum_s\rho_{sj}[Y_j]X_s+\sum_{r\neq j}\psi[Y_j,Y_r]}}{\sum_{\ell=1}^{L_j}\mathrm{e}^{\psi_{jj}[\ell,\ell]+\sum_s\rho_{sj}[\ell]X_s+\sum_{r\neq j}\psi[\ell,Y_r]}} \tag{9.30}$$

由此，伪对数似然概率可以定义为

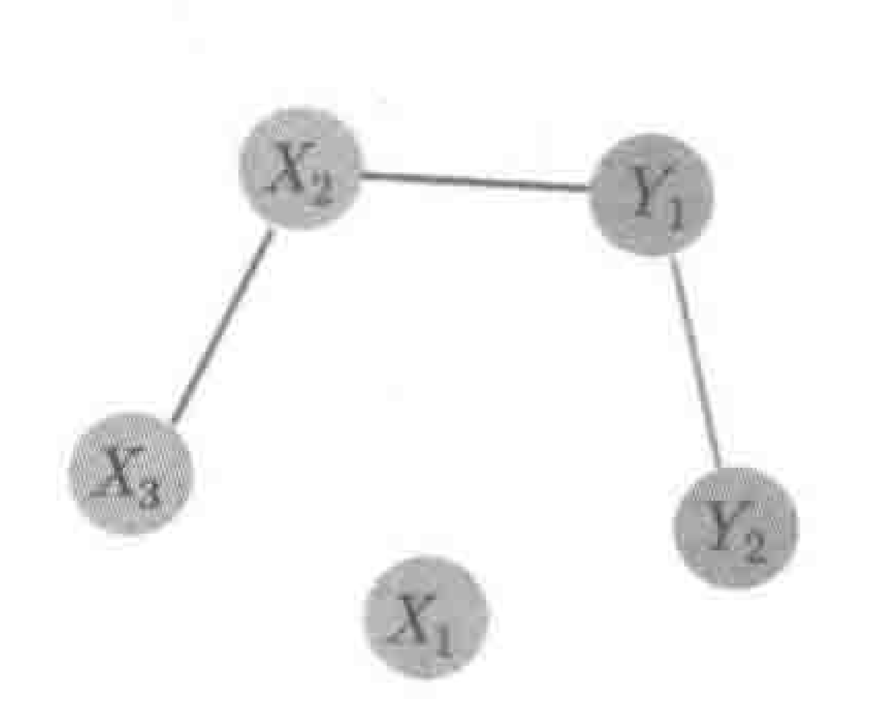
$$\ell^p(\boldsymbol{\Omega};\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{N}\left[\sum_{S=1}^{P}\log\mathbb{P}(x_{is}|x_{i\setminus\{s\}},y_i;\boldsymbol{\Omega})\sum_{j=1}^{q}\log\mathbb{P}(y_{ij}|x_i,y_{i\setminus\{j\}};\boldsymbol{\Omega})\right] \tag{9.31}$$

可以证明，式 (9.31) 是 $\boldsymbol{\Omega}$ 的凹函数。注意，各个参数出现了两次：下标中的一项对应于响应变量时出现一次，下标指向一个条件变量时出现第二次。

边参数有三类：标量 θ_{st}、向量 ρ_{sj}、矩阵 $\boldsymbol{\psi}_{jr}$。图 9-9 用一个小例子来说明这一点。Lee and Hastie（2014）用组 lasso 来选择这些不同的参数类型，他们提出优化惩罚伪对数似然概率

$$\ell^p(\boldsymbol{\Omega};\boldsymbol{X},\boldsymbol{Y})-\lambda\left(\sum_{s=1}^{p}\sum_{t=1}^{s-1}|\theta_{st}|+\sum_{s=1}^{p}\sum_{j=1}^{q}\|\boldsymbol{\rho}_{sj}\|_2+\sum_{j=1}^{q}\sum_{r=1}^{j-1}\|\boldsymbol{\psi}_{jr}\|_2\right)\tag{9.32}$$

其中，$\boldsymbol{\Omega}$ 为参数。注意，并非所有的参数都被惩罚，特别是，$\boldsymbol{\Theta}$ 的对角项留了下来（同图 lasso 算法一样），如同节点–潜能 $\boldsymbol{\psi}_{jj}$ 的各项。这就对一些参数估计产生了有趣的约束，这将在习题 9.9 中探讨。



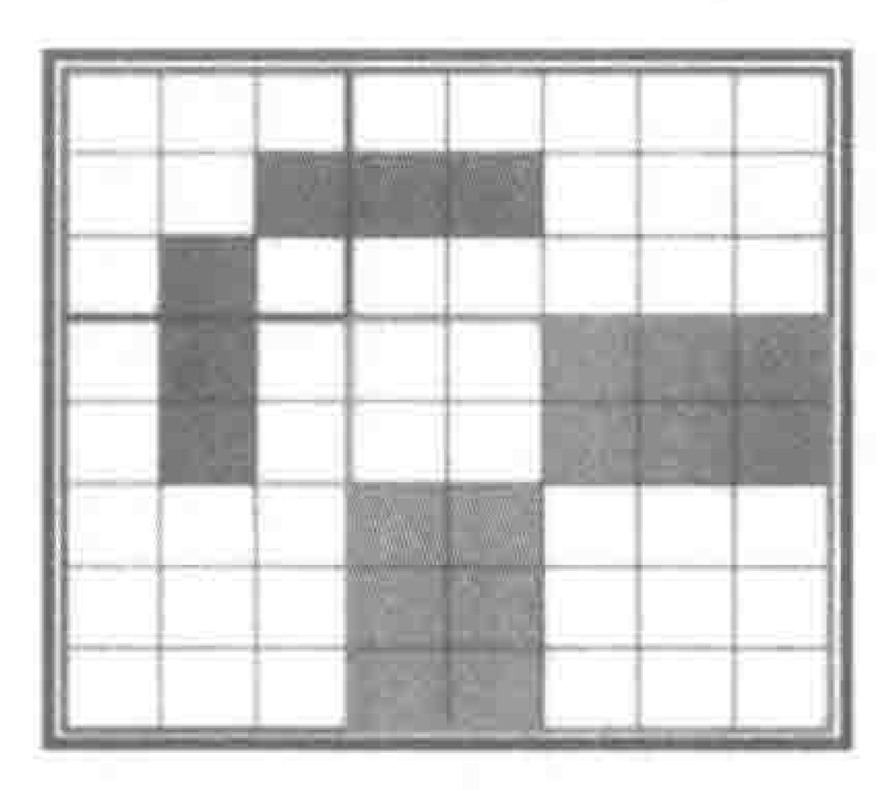


图 9-9　混合图模型，带有三个连续变量，两个离散变量。Y_1 有两个状态，Y_3 有三个。右边的图为和各边有关的参数集。组 lasso 将这些看做集合，整体包括或排除这些

如果只有连续变量，这就是一个高斯图模型下的惩罚伪似然概率。对全部二值变量，可以证明这等价于伊辛模型下的 lasso 惩罚伪似然概率（见 Hoefling and Tibshirani 2009 及习题 9.10）。

块坐标下降法在这里很有吸引力，因为其中各部分都有研究透彻的解。但是，参数是共享的，所以必须小心这种对称性。Lee and Hastie（2014）用了近似牛顿算法。

9.5　带隐变量的图模型

Chandrasekaran et al.（2012）提出了一种处理无向图模型的方法，其中一些变量是不可见的（即“隐藏”）。假设要对一个特定行业的股票价格建模，它们又极度依赖于能源的价格，后者在已掌握的数据中不可测。而股票价格的集中矩阵在已掌握的数据中看起来并不稀疏，但是如果依赖于能源价格，可能就会稀疏。

设所有可见和不可见的变量协方差矩阵为 $\boldsymbol{\Sigma}$。$\boldsymbol{\Sigma}$ 中对应可见变量的子块为 $\boldsymbol{\Sigma}_O$。设 $\boldsymbol{K}=\boldsymbol{\Sigma}^{-1}$ 是可见和隐藏变量集的集中矩阵，有子矩阵 $\boldsymbol{K}_O$、$\boldsymbol{K}_{O,H}$、$\boldsymbol{K}_{H,O}$

和 $\boldsymbol{K}_H$。这些子矩阵描绘了可见变量内、可见和隐藏变量之间，以及隐藏变量内各自的依赖关系。利用分块-求逆公式，可以将可见变量集中矩阵表示为

$$\tilde{\boldsymbol{K}}_O = \boldsymbol{\Sigma}_O^{-1} = \boldsymbol{K}_O - \boldsymbol{K}_{O,H}\boldsymbol{K}_H^{-1}\boldsymbol{K}_{H,O} \tag{9.33}$$

这里 $\boldsymbol{K}_O$ 是给定隐藏变量下可见变量的条件统计量的隐藏矩阵。现在 $\widetilde{\boldsymbol{K}}_O$ 可能不稀疏，但是如果图模型中所有变量的边较少，那么 $\boldsymbol{K}_O$ 将会稀疏。

由式（9.33）得到启发，设 $\boldsymbol{K}_O = \boldsymbol{\Theta}$，于是有

$$\tilde{\boldsymbol{K}}_O = \boldsymbol{\Theta} - \boldsymbol{L} \tag{9.34}$$

其中 $\boldsymbol{L}$ 的秩比较小，最多为隐藏变量的数目。接下来，在集合 $\{\boldsymbol{\Theta} - \boldsymbol{L} \succ \boldsymbol{0}, \boldsymbol{L} \succeq \boldsymbol{0}\}$ 上求解

$$\underset{\boldsymbol{\Theta},\boldsymbol{L}}{\text{minimize}} \{\text{trace}[\boldsymbol{S}(\boldsymbol{\Theta} - \boldsymbol{L})] - \log[\det(\boldsymbol{\Theta} - \boldsymbol{L})] + \lambda \|\boldsymbol{\Theta}\|_1 + \text{trace}(\boldsymbol{L})\} \tag{9.35}$$

同图 lasso 一样，这是一个凸问题。这点和 Mazumder et al.（2010）以及 Chandrasekaran et al.（2011）中讨论的“稀疏低秩”有关。Ma, Xue and Zou（2013）提出了一阶交替方向乘子（ADMM）法来解决该问题，并将它们和二阶方法比较，详见习题 9.11。

参考文献注释

Whittaker（1990）、Lauritzen （1996）、Cox and Wermuth（1996）、Edwards（2000）、Pearl（2000）, Anderson（2003）以及 Koller and Friedman（2009）详细地讨论了图模型。

Hammersley-Clifford 理论第一次是在 Hammersley and Clifford（1971）未出版的笔记中提出的。Besag（1974）和 Grimmett（1973）给出了独立证明，后者运用了 Möebius 倒置公式。一些历史讨论和结果的由来见 Clifford（1990）。

Welsh (1993) 讨论了估计一般离散图模型配分函数的难解之处。对于特殊结构的图模型，累积函数的精确求解可能是多项式时间。比如，树宽度较小的图可以运用联合树算法（Lauritzen and Spiegelhalter 1988，Lauritzen 1996）和一些平面模型（Kastelyn 1963, Fisher 1966）。其他一些例子可以用快速混合马尔可夫链来获得累积函数的较好近似（Jerrum and Sinclair 1993，Jerrum and Sinclair 1996）。另一种补充的方法是运用变分法，给出累积生成函数的近似（Wainwright and Jordan 2008）。例子包括平均场算法、和积或者置信传播算法、期望传播，以及其他多种凸松弛方法。对于一些图，尤其是“局部类树”图，有各种各样的渐近精确结果（M’ezard and Montanari 2008）。

高斯图模型可以用来对基因表达数据（Dobra, Hans, Jones, Nevins, Yao and West 2004）和其他基因、蛋白质实验建模。伊辛模型（9.4）由伊辛在统计物理学中首次提出（1925）。在近期的研究中，伊辛及其相关模型用于二值图像的简化模型（Geman and Geman 1984，Greig，Porteous and Seheuly 1989，Winkler 1995）、政客投票行为（Banerjee et al. 2008），引用网络分析（Zhao, Levina and Zhu 2011）。

本章讨论的关于无向图模型的一些方法可用于帮助构建一些复杂模型（用于有向图模型），见 Schmidt, Niculescu-Mizil and Murphy（2007）。Vandenberghe et al. （1998）的文章引入了带约束的行列式最大化问题，高斯 MLE（带或者不带正则化）是这类问题的一个特例。Yuan and Lin（2006b）提出了对方差–选择问题用 ℓ_1 正则化及高斯（log 行列式）似然概率构建模型，并用内点法（Vandenberghe et al. 1998）求解。d'Aspremont et al.（2008）和 Friedman et al.（2008）提出了一种快速坐标下降算法，基于一系列子问题求解图 lasso 式（9.13）。Mazumder and Hastie（2012）在这些算法上提出了改进，使其有更好的收敛特性。Witten et al.（2011）和 Mazumder and Hastie（2012）说明了如何在计算图 lasso 解时利用 $\boldsymbol{S}$ 中的块对角结构。Rothman, Bickel, Levina and Zhu（2008）确定了 Frobenius 范数中估计的一致性，而 Ravikumar et al.（2011）给出了关于模型选择一致性及算子范数速率的一些结果。他们还证明了图 9-5 所示的算子范数界（9.21）。

伪似然概率的思想有一定的历史，可以追溯到 Besag（1975）的开创性工作。Meinshausen and Bühlmann（2006）第一次提出及发展了针对高斯图模型的基于 lasso 的近邻选择法，并且在高维尺度变换下推导出了一致性结果，也可参考 Zhao and Yu（2006）和 Wainwright（2009）中静态图上的相关结果。Zhou, Lafferty and Wasserman（2008）考虑了跟踪高斯图模型的时变序列问题。

Ravikumar, Wainwright and Lafferty（2010）对于离散二值图模型上的模型选择问题提出了构建 ℓ_1 正则化逻辑斯蒂回归，并且证明了尺度 $N = \Omega\left(d^3 \log p\right)$ 下有模型选择一致性。Bento and Montanari（2009）随后的分析证明了对应低于相变的伊辛模型，这个尺度为 $N = \Omega\left(d^2 \log p\right)$。Koh et al.（2007）提出了一种适用于大型 ℓ_1 正则化逻辑斯蒂回归的内点算法。不用在每个节点求解单个逻辑斯蒂回归，Hoefling and Tibshirani（2009）提出了最小化 ℓ_1 正则化伪似然概率，并且推导了求解的有效算法，类似论述可参考 Friedman et al.（2010a）。Santhanam and Wainwright（2008）推导了伊辛选择模型中的信息论下界，证明了如果 $N = \mathcal{O}\left(d^2 \log p\right)$，则多数情况无从求解。这说明即使在常数因子下，近邻法也是最优方法。

Cheng, Levina and Zhu（2013）和 Lee and Hastie（2014）讨论了混合图模型，包括连续变量和离散变量。Kalischand Bühlmann（2007）证明了 PC 算法的变体可以用于有向图的高维模型选择。

另一种图模型是*协方差图*或者*关联网络*，如果变量间的协方差矩阵（而非部分

协方差）非零，则对应顶点有双向边链接。这在几何学中十分流行，例子可见 Butte, Tamayo, Slonim, Golub and Kohane（2000）。这些模型的负对数似然函数是非凸的，使得计算更加复杂（Chaudhuri, Drton and Richardson 2007）。Bien and Tibshirani（2011）和 Wang（2014）取得了关于此问题的最新进展。后者提出了一个类似于图 lasso 算法情形的块坐标下降算法。Bickel and Levina（2008）和 El Karoui（2008）进行了大型协方差矩阵估计的一些理论研究。

习 题

习题 9.1 多维高斯分布中最常见的参数化是关于均值向量 $\mu \in \mathbb{R}^p$ 和协方差矩阵 $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ 的。设分布非退化（即 $\boldsymbol{\Sigma}$ 严格正定），求证分解式（9.8）中的典型参数 $(\gamma, \boldsymbol{\Theta}) \in \mathbb{R}^p \times \mathcal{S}_+^p$ 与下式相关：

$$\mu = -\boldsymbol{\Theta}^{-1}\gamma \text{和} \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1} \tag{9.36}$$

习题 9.2 设 $\{x_1, \ldots, x_N\}$ 为高斯图模型上得到的独立同分布样本，设 $\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{X}) = \dfrac{1}{N}\sum_{i=1}^N \log \mathbb{P}_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)$ 表示尺度调整后样本的对数似然概率。

(a) 求证：

$$L(\boldsymbol{\Theta}; \boldsymbol{X}) = \log\det \boldsymbol{\Theta} - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Theta}) + C$$

其中 $\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^N x_i x_i^{\mathrm{T}}$ 是经验协方差矩阵，C 是一个独立于 $\boldsymbol{\Theta}$ 的常数。

(b) 求证：函数 $f(\boldsymbol{\Theta}) = -\log\det\boldsymbol{\Theta}$ 在正定矩阵锥形上是一个严格凸函数，且对于任意 $\boldsymbol{\Theta} \in \mathcal{S}_+^p$，有 $\nabla f(\boldsymbol{\Theta}) = \boldsymbol{\Theta}^{-1}$。

(c)（未正则化）高斯 MLE 由下式给出：

$$\hat{\boldsymbol{\Theta}} \in \underset{\boldsymbol{\Theta} \in \mathcal{S}_+^p}{\arg\max}\{\log\det\boldsymbol{\Theta} - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Theta})\}$$

假设已经达到最大值，求证 $\widehat{\boldsymbol{\Theta}} = \boldsymbol{S}^{-1}$。讨论当 $N < P$ 时会有什么变化。

(d) 现在考虑用图 lasso 式（9.13），这是基于增强缩放的 log 似然函数得到的，该似然函数带有 ℓ_1 正则化。推导 Karush–Kuhn–Tucker 公式，使任意原始–最优对 $(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{W}}) \in \mathcal{S}_+^p \times \mathbb{R}^{p \times p}$ 必须满足该公式。

(e) 推导图 lasso 有关的对偶问题。能否将你的结果推广到带任意 $\ell_q(q \in [1, \infty])$ 范数的正则化？

习题 9.3 求证：如果 $\boldsymbol{S}$ 正定，λ=0 下图 lasso 算法可以计算得到 $\widehat{\boldsymbol{\Theta}} = \boldsymbol{S}^{-1}$。

习题 9.4　本习题利用联合高斯随机向量的特性，保证协方差选择问题中基于近邻的 lasso 方法的 Fisher 一致性。设 $\{X_1, X_2, \ldots, X_p\}$ 是一个零均值联合高斯随机向量，其正定协方差矩阵为 $\boldsymbol{\Sigma}$。设 $T = \{2, 3, \ldots, p\}$，解答下列关于条件随机变量 $Z = (X_1 \,|\, X_T)$ 的问题。

(a) 求证: 存在向量 $\theta \in \mathbb{R}^{p-1}$ 使得

$$Z = \theta^{\mathrm{T}} X_T + W$$

其中 W 是一个独立于 X_T 的零均值高斯变量。**提示**: 考虑给定 X_T 下的 X_1 的最佳线性预测子。

(b) 求证: $\theta = \boldsymbol{\Sigma}_{TT}^{-1} \boldsymbol{\Sigma}_{T1}$，其中 $\boldsymbol{\Sigma}_{T1} \in \mathbb{R}^{p-1}$ 是 X_T 和 X_1 之间协方差的向量。

(c) 求证: 当且仅当 $j \notin \mathcal{N}(1)$，有 $\theta_j = 0$。**提示**: 可应用如下基本定理: 设 $\boldsymbol{A}$ 是一个不可逆矩阵，给定块分割形式

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$$

然后设 $\boldsymbol{B} = \boldsymbol{A}^{-1}$，有 $\boldsymbol{B}_{12} = \boldsymbol{A}_{11}^{-1} \boldsymbol{A}_{12} \left[\boldsymbol{A}_{21} \boldsymbol{A}_{11}^{-1} \boldsymbol{A}_{12} - \boldsymbol{A}_{22} \right]^{-1}$（Horn and Johnson 1985）。

习题 9.5　思考伊辛模型选择上基于近邻的似然概率方法，并解答如下问题。

(a) 推导条件分布 $\mathbb{P}\left(x_s | x_{V \setminus \{s\}}; \theta\right)$，展示近邻预测如何降为逻辑斯蒂回归。

(b) 验证该方法是 Fisher 一致的，即真实条件分布为群体最小化。

习题 9.6　证明，如何针对式（9.14）一次一行一列地求解 $\boldsymbol{\Theta}$ 及其逆 $\boldsymbol{W} = \boldsymbol{\Theta}^{-1}$。为了简洁，这里首先关注最后一行一列。式（9.14）中的右上方块可以写成

$$\boldsymbol{w}_{12} - \boldsymbol{s}_{12} - \lambda \cdot \operatorname{sgn}(\boldsymbol{\theta}_{12}) = 0 \tag{9.37}$$

这里将矩阵分为两部分，一部分是前面 $p-1$ 行和列，另一部分是第 p 行和列。$\boldsymbol{W}$ 及其逆 $\boldsymbol{\Theta}$ 也以相似的方式分割

$$\begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}^{\mathrm{T}} & w_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}^{\mathrm{T}} & \theta_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{pmatrix} \tag{9.38}$$

求证:

$$\boldsymbol{w}_{12} = -\boldsymbol{W}_{11} \boldsymbol{\theta}_{12} / \theta_{22} \tag{9.39}$$

$$= \boldsymbol{W}_{11} \beta \tag{9.40}$$

其中 $\beta = -\boldsymbol{\theta}_{12}/\theta_{22}$。这是从一个矩阵的分割逆的逆公式中得到的（Horn and Johnson 1985）。将式（9.40）代入式（9.37）得到

$$\boldsymbol{W}_{11}\beta - \boldsymbol{s}_{12} + \lambda \cdot \mathrm{sgn}(\beta) = 0 \tag{9.41}$$

习题 9.7 如同式（9.38）中分割后，写出各个矩阵的分块逆表达式，表达式用其他矩阵表示。证明，既然 $\boldsymbol{W}_{11}$ 依赖于 $\boldsymbol{\theta}_{12}$，我们也不必如同图 lasso 算法 9.1 中一样固定住 $\boldsymbol{W}_{11}$。

(a) 求证：式（9.37）可以写成

$$\boldsymbol{\Theta}_{11}^{-1}\boldsymbol{\theta}_{12}w_{22} + \boldsymbol{s}_{12} + \lambda\mathrm{sgn}(\boldsymbol{\theta}_{12}) = \mathbf{0} \tag{9.42}$$

(b) 证明如何用解 $\boldsymbol{\theta}_{12}$ 来迭代 $O(p^2)$ 算子中的 $\boldsymbol{W}$ 和 $\widehat{\boldsymbol{\Theta}}$。

(c) 证明如何经过 $O(p^2)$ 次操作移至一个新的方程组。

(d) 本习题已经推导出了一个原始的图 lasso 算法。请以算法形式记录下来，参见算法 9.1。

（Mazumder and Hastie 2012）

习题 9.8 推导混合图模型中的条件分布式（9.29）和式（9.30）。

习题 9.9 仔细检查成对马尔可夫随机场模型（9.28）会发现在离散势 ρ_{sj} 和 ψ_{jr} 上的过参数化。本习题证明这种混淆可以通过惩罚伪似然函数中的二次惩罚来解决，以回归和 ANOVA 模型中常见的"总和为零"为约束形式。

思考惩罚伪 log 似然函数（9.32），$\lambda > 0$，解答如下问题。

(a) r_s 没有被惩罚，求证：解 $\widehat{\rho}_{sj}$ 对任意 s 和 j 满足

$$\sum_{\ell=1}^{L_j} \hat{\rho}_{sj}[\ell] = 0$$

(b)（对角）矩阵 $\boldsymbol{\psi}_{jj}$ 没有被惩罚，求证：解 $\widehat{\boldsymbol{\psi}}_{jr}$ 对任意 $j \neq r$ 满足

$$\sum_{\ell=1}^{L_j} \hat{\boldsymbol{\psi}}_{jr}[\ell, m] = 0 \quad m = 1, \ldots, L_r \tag{9.43}$$

$$\sum_{m=1}^{L_r} \hat{\boldsymbol{\psi}}_{jr}[\ell, m] = 0 \quad \ell = 1, \ldots, L_j \tag{9.44}$$

习题 9.10 思考成对马尔可夫随机场模型，设模型中只含二值离散变量。这似乎与伊辛模型不同，因为每条边有 4 个参数。用习题 9.9 的结论证明，有了式（9.32）中的二次约束，它正好等于伊辛模型下的 lasso 惩罚伪对数似然概率。

习题 9.11　考虑允许隐藏变量的图模型的目标函数（9.35）。定义一个新变量 $\boldsymbol{R} = \boldsymbol{\Theta} - \boldsymbol{L}$，详细推导先后在 $\boldsymbol{R}$、$\boldsymbol{\Theta}$、$\boldsymbol{L}$ 和 $\boldsymbol{\Gamma}$ 上（Ma et al. 2013）用扩展拉格朗日公式求解式（9.35）时采用的 ADMM 算法。扩展拉格朗日公式为

$$\begin{aligned} L_\mu(\boldsymbol{R}, \boldsymbol{\Theta}_0, \boldsymbol{L}, \boldsymbol{\Gamma}) = & \operatorname{trace}(\boldsymbol{S}\boldsymbol{R}) - \log\det \boldsymbol{R} + \lambda \|\boldsymbol{\Theta}\|_1 \\ & + \beta \cdot \operatorname{trace}(\boldsymbol{L}) + I(\boldsymbol{L} \succeq \mathbf{0}) - \operatorname{trace}[\boldsymbol{\Gamma}(\boldsymbol{R} - \boldsymbol{\Theta} + \boldsymbol{L})] \\ & + \frac{1}{2\mu} \|\boldsymbol{R} - \boldsymbol{\Theta} + \boldsymbol{L}\|_F^2 \end{aligned}$$

第 10 章　信号近似与压缩感知

10.1　引　　言

本章会介绍基于 ℓ_1 松弛的信号恢复与近似问题。重点介绍稀疏性在信号表示和近似中的作用，以及如何利用 ℓ_1 方法得到所求解问题（如信号去噪、压缩、近似）的稀疏性。这里首先举例说明：当由适当的基来表示时（比如由小波和多尺度变换表示），很多信号“天生”就是稀疏的。下面会举例说明如何采用正交基压缩和去噪得到稀疏性。接下来会讨论过完备基（overcomplete base）的信号近似问题，以及 ℓ_1 松弛寻找近似最优的作用。最后介绍恢复稀疏信号时所用的压缩感知方法。这里需要将两个概念结合起来：通过随机投影得到信号的度量，以及针对重构的 lasso 问题的求解方法。

10.2　信号与稀疏表示

我们首先介绍一下稀疏表示在信号处理中的背景知识。需要注意，这里所说的“信号”是广义的，包括数据，比如海平面数据、地震记录、医疗时间序列、音频记录、摄影图像、视频数据和金融数据。在这些情况下，信号都可用向量 $\theta^* \in \mathbb{R}^p$ 来表示。（图像等二维信号可以看成向量化的形式）。

10.2.1　正交基

在处理信号时，用不同的基来表示信号通常会很有用，这些基可以采用傅里叶表示和多尺度表示（比如小波）。在时间序列中，傅里叶表示对提取周期性结构很实用。这样的表示形式是一组向量 $\{\psi_j\}_{j=1}^p$，这组向量为 $\mathbb{R}^p$ 中的正交基。定义一个 $p \times p$ 的矩阵 $\boldsymbol{\varPsi} := [\psi_1 \quad \psi_2 \quad \ldots \quad \psi_p]$，则由正交条件可知 $\boldsymbol{\varPsi}^{\mathrm{T}}\boldsymbol{\varPsi} = \boldsymbol{I}_{p\times p}$。给定一组正交基，则任意的信号 $\theta^* \in \mathbb{R}^p$ 可表示为

$$\theta^* := \sum_{j=1}^{p} \beta_j^* \psi_j \tag{10.1}$$

其中，第 j 个基系数 $\beta_j^* := \langle \theta^*, \boldsymbol{\psi}_j \rangle = \sum_{i=1}^p \theta_i^* \psi_{ij}$ 是将信号投影到第 j 个基 $\boldsymbol{\psi}_j$ 而得到的。因此，可将信号 $\theta^* \in \mathbb{R}^p$ 的基系数 $\beta^* \in \mathbb{R}^p$ 写成矩阵与向量相乘的形式：$\beta^* = \boldsymbol{\varPsi}^{\mathrm{T}}\theta^*$。

下面举一个简单的例子来进行说明。设有矩阵

$$\Psi := \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{-1}{\sqrt{2}} \end{bmatrix} \tag{10.2}$$

这是一个正交矩阵，即 $\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \boldsymbol{I}_{4\times 4}$，用它对长度为 $p=4$ 的信号进行相应的两级哈尔变换。对任意给定的信号 $\theta^* \in \mathbb{R}^4$，哈尔基系数 $\beta^* = \boldsymbol{\Psi}^{\mathrm{T}}\theta^*$。第一个系数 $\beta_1^* = \langle \boldsymbol{\psi}_1, \theta^* \rangle = \frac{1}{2}\sum_{j=1}^{4} \theta_j^*$ 是平均信号重新调整后的版本。第二列 ψ_2 是对整个信号做差分操作的结果，而第三列和第四列是对每半个信号做局部差分操作的结果。哈尔变换是最简单的小波变换。

一个重要的事实是，许多种信号在标准的基下并不稀疏，而用特殊的正交基来表示时才会变得稀疏。图 10-1 为一些医学上的时序数据示意图。左上图是从一名患者身上采集到的 $p=128$ 个时间点的动脉血压信号 θ^*，可以看出信号 θ^* 本身并不全都稀疏。右图为信号在哈尔基下的系数 $\beta^* = \boldsymbol{\Psi}^{\mathrm{T}}\theta^*$。注意：相比之下这是相对稀疏的。最后，左下图给出了原始信号的重构 $\hat{\theta}$，该重构丢掉了一半的哈尔系数。这虽然不是一个完美的重构，却得到了时间序列数据的主要特征。

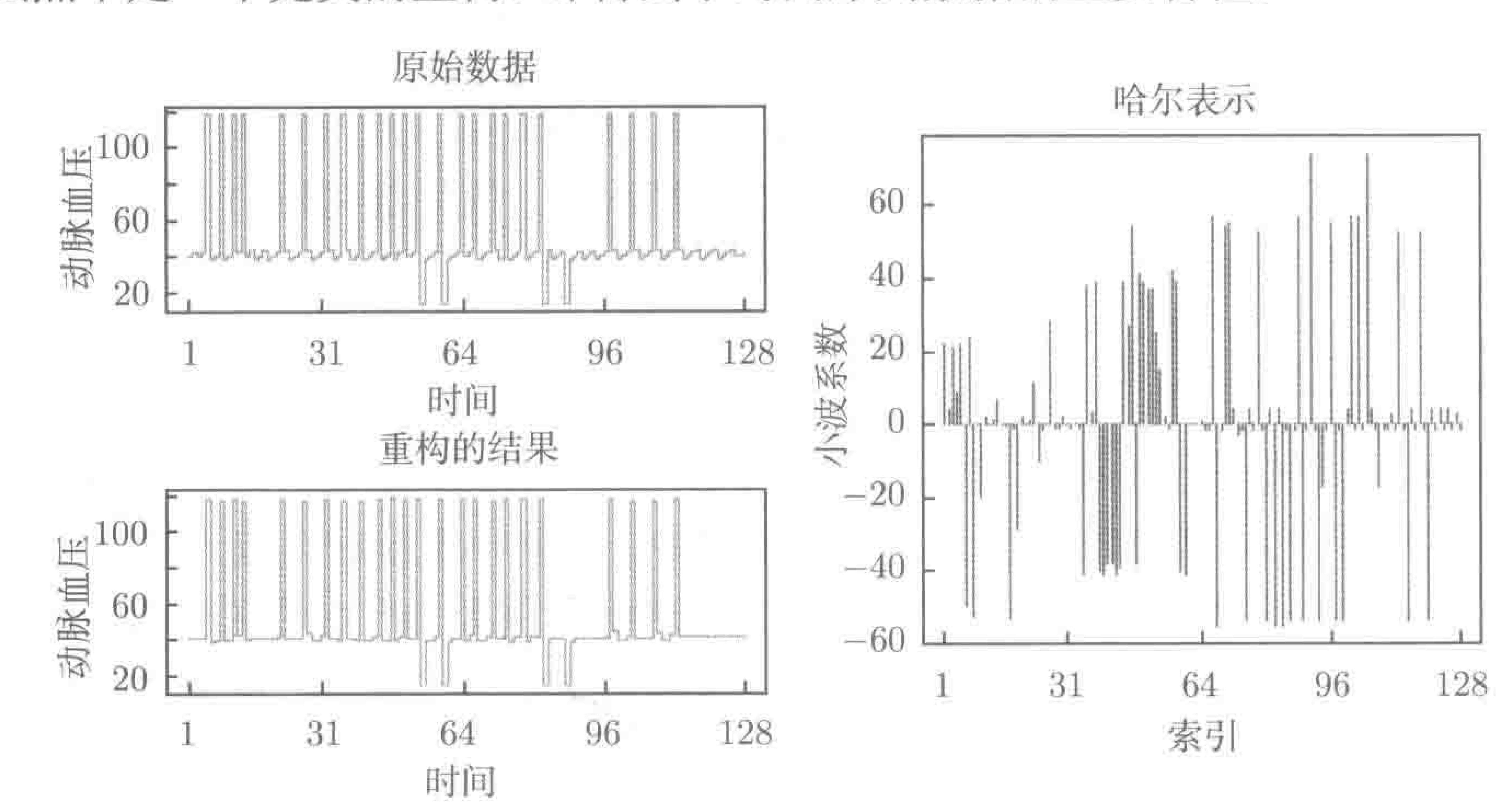


图 10-1　时序数据的稀疏示意图。左上图：有 $p=128$ 个时间点的动脉血压信号 θ^*。左下图：从哈尔基中（以绝对振幅）取最大的前 $k=64$ 个系数，重构 $\hat{\theta}^{128}$。右图：$\beta^* = \boldsymbol{\Psi}^{\mathrm{T}}\theta^*$ 在哈尔基下的系数

图 10-2 给出了这种稀疏现象的第二个示意图，图中有拍摄的图像和二维小波变换的结果。图 10-2a 为一艘船的图像，其大小为 512×512，我们可将这幅二维图

像当成一个大小为 $p = 512^2 = 262\,144$ 的向量。图 10-2b 为这幅图的一个特定二维小波形式，这可从图中的形状看出来，所设计的小波要在具体尺度下提取对角线方向上的结构。将图像的所有空间位置与该小波作内积（称为卷积过程），就会得到图像在所有空间位置上的小波系数。接下来对这些系数进行子采样（sub-sample），这取决于小波的尺度。然后用这些系数来重建图像。在多尺度（对本图采用了三级尺度）和方向（对本图采用了四个方向）下做同样的操作，就会得到多尺度金字塔，如图 10-2c 所示。尽管原始图像并不是稀疏信号，但用这样的多尺度基来表示时会变得非常稀疏，因为它的许多系数为零或非常接近零。为了阐明这种稀疏性，图 10-2d 展示了一些小波系数的直方图。直方图是对图像所有空间位置上的小波系数池化（pooling）后的结果。该直方图按对数尺度绘制，在零附近有很尖的峰值，这表明系数分布很稀疏。

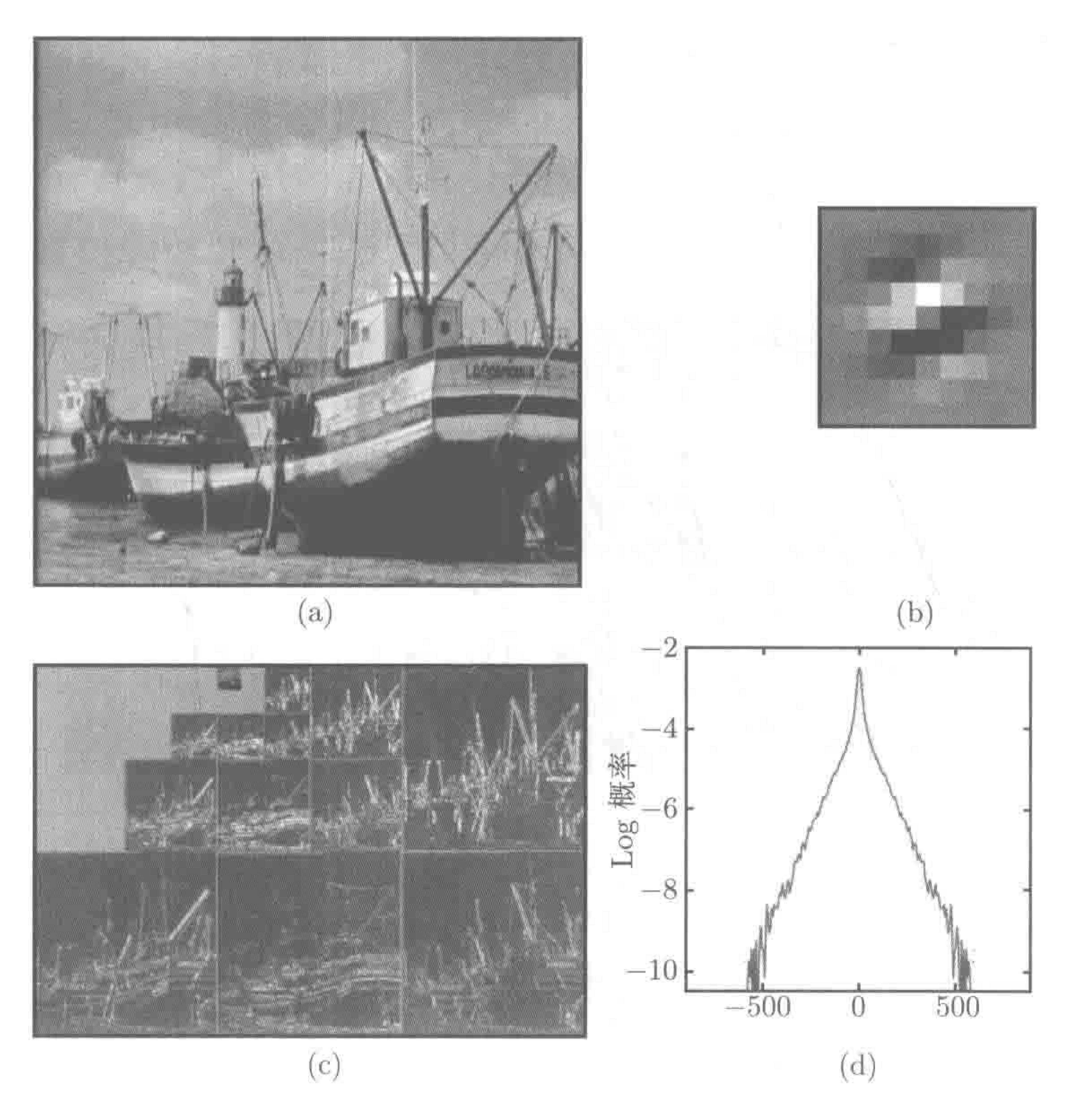


图 10-2 用小波基表示图像而得到的稀疏性。(a) 一艘船的图像。(b) 多尺度金字塔变换的基向量，这里用二维图像来展示这些基向量。(c) 用三级多尺度来表示这幅图像，每个尺度有四个不同方向。(d) 在某个尺度和方向上，小波系数的幅值对数直方图，对图像中的所有像素进行池化而得到。请注意，多数的系数都接近于零，绝对值相对较大的系数很少

10.2.2 用正交基逼近

信号压缩的目的是表示信号 $\theta^* \in \mathbb{R}^p$，这通常是一种近似手段，使用一定数量的系数，所用到的数量要比原始维数小得多（$k \ll p$）。在正交基的情形下，有一种方法仅用正交向量 $\{\psi_j\}_{j=1}^p$ 的稀疏子集就能完成表示。比如，用一个整数 $k \in (1,2,\ldots,p)$ 来控制近似精度，则重构形式为

$$\boldsymbol{\Psi}\beta = \sum_{j=1}^{p} \beta_j \psi_j, \quad \text{使得} \|\beta\|_0 := \sum\nolimits_{j=1}^{p} \mathbb{I}[\beta_j \neq 0] \leqslant k \tag{10.3}$$

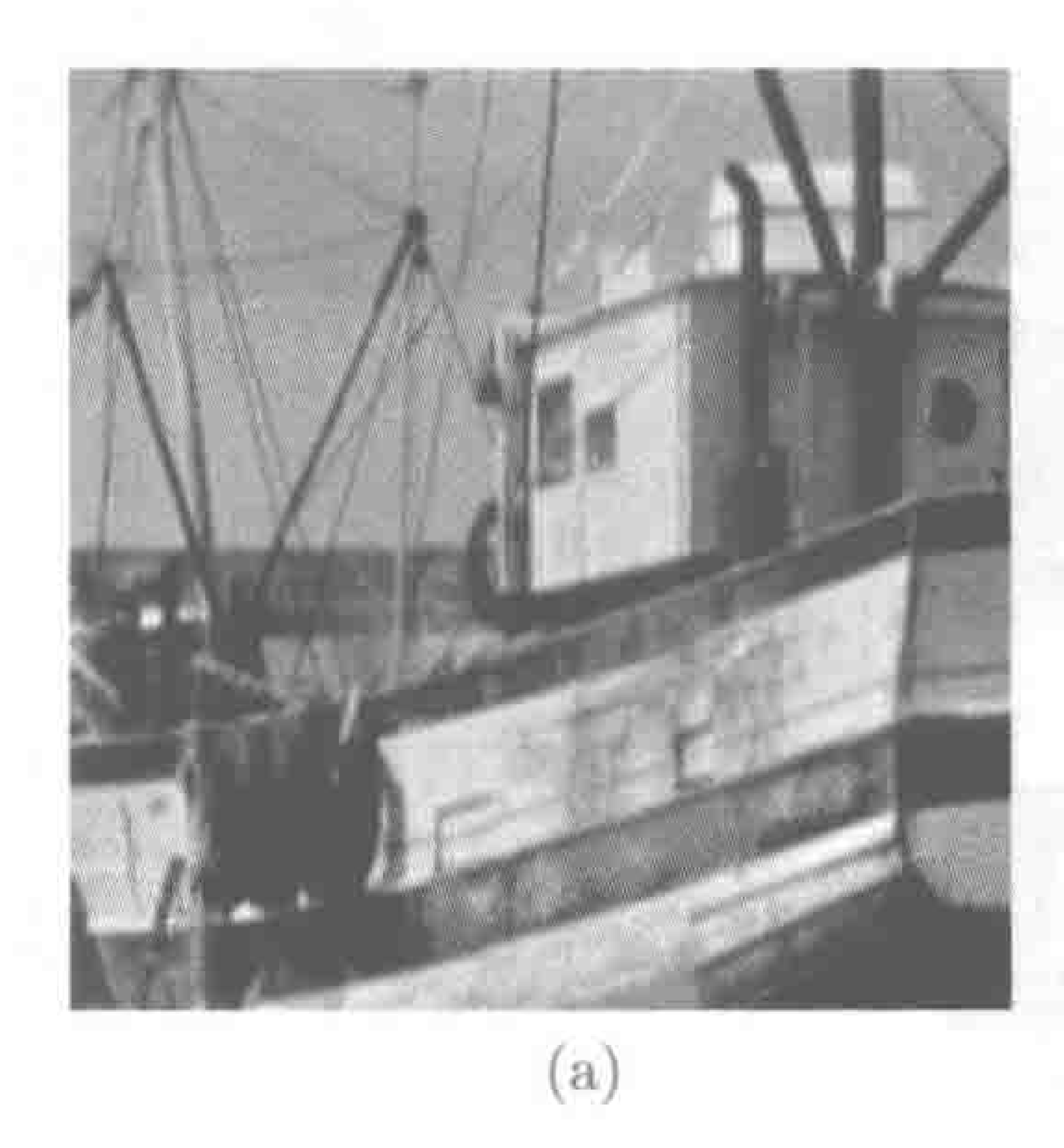
这里引入了 ℓ_0 范数，对向量 $\beta \in \mathbb{R}^p$ 中的非零元素进行计数。下面是这个问题的 k 稀疏近似

$$\hat{\beta}^k \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\|\theta^* - \boldsymbol{\Psi}^{\mathrm{T}}\beta\right\|_2^2, \quad \text{使得} \|\boldsymbol{\beta}\|_0 \leqslant k \tag{10.4}$$

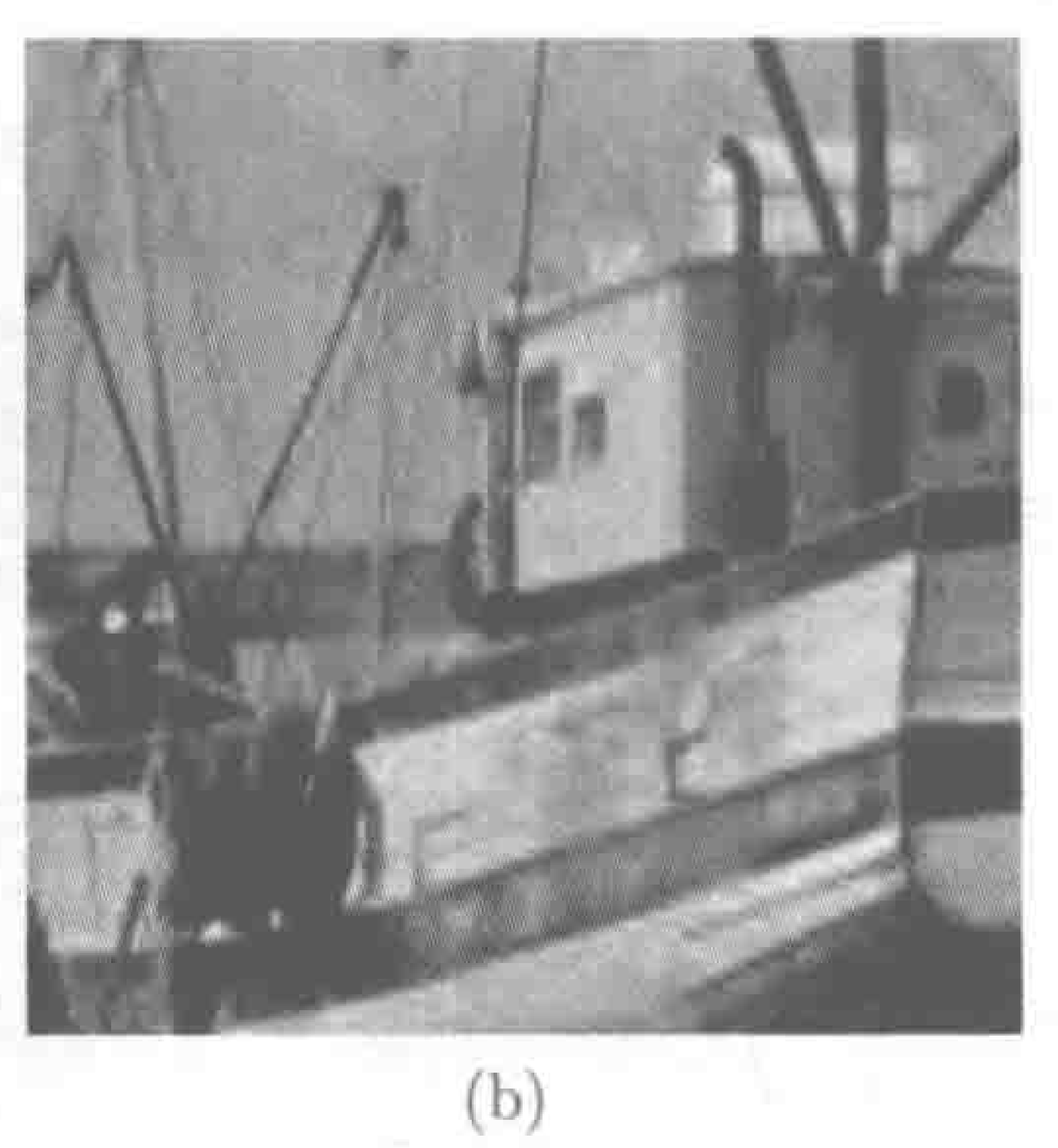
得到该问题的最优解 $\hat{\beta}^k$ 后，可重构为

$$\theta^k := \sum_{j=1}^{p} \hat{\beta}_j^k \psi_j \tag{10.5}$$

这里通过 k 个最佳系数来近似 θ^*，如图 10-3 所示。



(a)



(b)

图 10-3　基于小波阈值的图像压缩示意图。(a) 放大图 10-2 中“船”部分图像的结果。(b) 用绝对值最大的前 5% 的小波系数重构的结果。注意，重构结果只有相当少的失真，主要集中在图像的细尺度特征上

注意，由于有 ℓ_0 范数约束，式（10.4）是一个非凸的组合优化问题。尽管如此，仍有一种简单的方法可以求解该问题，即通过正交变换得到结构。假设按绝对值来对基的系数向量 $\beta^* \in \mathbb{R}^p$ 中的各元素进行排序，即

$$\left|\beta^*_{(1)}\right| \geqslant \left|\beta^*_{(2)}\right| \geqslant \ldots \geqslant \left|\beta^*_{(p)}\right| \tag{10.6}$$

对于任意给定的整数 $k \in (1,2,\ldots,p)$，则最优的 k 近似为

$$\hat{\theta}^k := \sum_{j=1}^{k} \beta^*_{(j)} \psi_{\sigma(j)} \tag{10.7}$$

其中，$\sigma(j)$ 表示排好序的基向量中的第 j 元素。也就是说，获得的基向量与最大（按绝对值）的 k 个系数有关。

总之，可以用正交基按如下算法计算最优的 k 项近似：

(1) 计算基系数 $\beta^*_j = \langle \theta^*, \psi_j \rangle$，$j = 1, 2, \ldots, p$，通过矩阵与向量相乘得到：$\beta^* = \boldsymbol{\Psi}^{\mathrm{T}} \theta^*$；

(2) 按系数绝对值排序，如式（10.6），取前 k 个系数；

(3) 计算最佳 k 项近似 $\hat{\theta}^k$，如式（10.7）。

对任意的正交基础，该过程的计算复杂度至多为 $\mathcal{O}(p^2)$，第 (2) 步中排序的复杂度为 $\mathcal{O}(p \log p)$，主要源于第 (1) 步对基的系数的计算复杂度。许多正交表示（比如傅里叶基和离散小波）有一个很好的性质，即计算基系数的时间复杂度为 $\mathcal{O}(p \log p)$。

如前所述，图 10-1 展示了用哈尔小波基来近似信号的示意图，左下图显示了近似信号 $\hat{\theta}^{64}$，即仅保留一半的哈尔小波系数（$k/p = 64/126 = 0.5$）。

10.2.3 用过完备基来重构

正交基虽然在很多方面都很有用，但也有缺点。具体而言，其类信号只能用某种正交基来进行稀疏表示。例如，傅里叶基非常适合重构有全局周期结构的信号，但具有局部化能力的哈尔基则不能很好捕获这种结构。哈尔基擅长表示不连续的情形（step discontinuity），而傅里叶基对这种情形不会得到稀疏表示。

基于这样的直觉，若信号在某种意义上是“简单”的，则可相对直接地构造，但传统的正交基下不具有稀疏性。图 10-4a 显示了信号 $\theta^* \in \mathbb{R}^{128}$，它由一些全局周期性成分和一些快速（几乎不连续）的变化混合而成。图 10-4b 中的哈尔系数 $\beta^* = \boldsymbol{\Psi}^{\mathrm{T}} \theta^*$ 比较稠密，因为许多基向量会用来重构信号的全局周期部分。类似地，图 10-4c 通过离散余弦基（一种傅里叶表示）得到的 $\alpha^* = \boldsymbol{\Phi}^{\mathrm{T}} \theta^*$ 也相对稠密。由于缺乏稀疏性，单个基不能对原始信号进行很好的稀疏近似。

但若同时允许用两个基的向量子集来进行重构，可能会得到更高精度，甚至可能得到完全稀疏的近似。为了让这个问题更精确，可用给定的两组正交基 $\{\psi_j\}_{j=1}^p$ 和 $\{\phi_j\}_{j=1}^p$ 得到重构形式

$$\underbrace{\sum_{j=1}^{p} \alpha_j \phi_j}_{\boldsymbol{\Phi}\alpha} + \underbrace{\sum_{j=1}^{p} \beta_j \psi_j}_{\boldsymbol{\Psi}\beta}, \quad \text{使得 } \|\alpha\|_0 + \|\beta\|_0 \leqslant k \tag{10.8}$$

相关的优化问题为

$$\underset{(\alpha,\beta) \in \mathbb{R}^p \times \mathbb{R}^p}{\text{minmize}} \|\theta^* - \boldsymbol{\Phi}\alpha - \boldsymbol{\Psi}\beta\|_2^2, \quad \text{使得 } \|\alpha\|_0 + \|\beta\|_0 \leqslant k \tag{10.9}$$

尽管上面的公式表面上与前面的 k 项近似问题（10.5）相似，但实际上优化问题（10.9）很难求解。与前面的不同在于，现在这种情形采用了由两个基 $\boldsymbol{\Phi}$ 和 $\boldsymbol{\Psi}$ 构成的过完备基来表示问题。

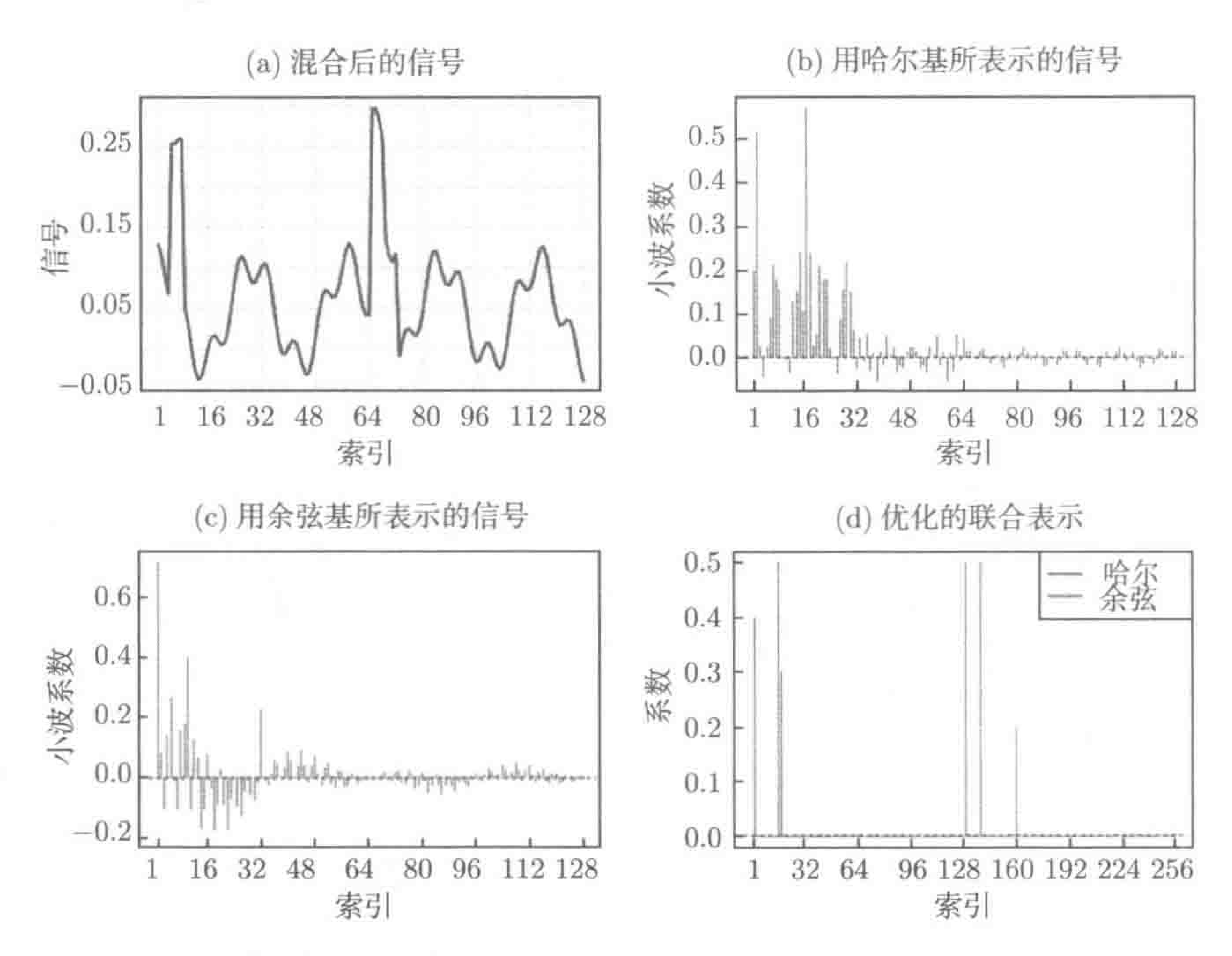


图 10-4　(a) 原始信号 $\theta^* \in \mathbb{R}^p$，其中 $p=128$。(b) 用哈尔基来表示 $\boldsymbol{\Psi}^{\mathrm{T}}\theta^*$。(c) 用离散余弦基表示 $\boldsymbol{\Phi}^{\mathrm{T}}\theta^*$。(d) 系数 $(\hat{\alpha},\hat{\beta}) \in \mathbb{R}^p \times \mathbb{R}^p$ 是优化后的联合稀疏表示，它可能通过基追踪线性规划 [见式（10.11）] 来求解

然而，通过对 ℓ_0 范数的松弛可以得到凸规划问题

$$\underset{(\alpha,\beta)\in\mathbb{R}^p\times\mathbb{R}^p}{\text{minmize}} \|\theta^* - \boldsymbol{\Phi}\alpha - \boldsymbol{\Psi}\beta\|_2^2, \quad 使得\ \|\alpha\|_1 + \|\beta\|_1 \leqslant R \tag{10.10}$$

其中，$R>0$ 是用户定义的半径。这个目标函数就是一个 lasso 的约束版本，也称为松弛的基追踪规划（basis-pursuit program）。为了得到一个好的重构，也可考虑更简单的问题

$$\underset{(\alpha,\beta)\in\mathbb{R}^p\times\mathbb{R}^p}{\text{minmize}} \|\alpha\|_1 + \|\beta\|_1, \quad 使得\ \theta^* = [\boldsymbol{\Phi} \quad \boldsymbol{\Psi}]\begin{bmatrix}\alpha \\ \beta\end{bmatrix} \tag{10.11}$$

这是一个线性规划问题，通常也称为基追踪线性规划。

回到图 10-4 所讨论的例子，图 10-4d 展示了由基追踪线性规划（10.11）得到的最优系数 $(\hat{\alpha},\hat{\beta}) \in \mathbb{R}^p \times \mathbb{R}^p$。由此可见，图 10-4a 的原始信号可通过极其稀疏的组

合（只有 6 个非零系数）来产生，过完备基可由哈尔基和离散余弦基来得到。其实，这是信号最可能的稀疏表示。因此，在这种情况下，求解基追踪线性规划（10.11）等价于求解带 ℓ_0 约束的问题（10.9）。

当然，读者可能会质疑这种现象的普遍性 —— 求解基追踪线性规划什么时候与求解带 ℓ_0 约束的式（10.9）等价？事实证明，这个问题的答案取决于两组基 $\boldsymbol{\Phi}$ 和 $\boldsymbol{\Psi}$ 的不相关（incoherence）程度。这方面的内容会在 10.4 节进行详细介绍。

10.3 随机投影与近似

上一节讨论了如何通过计算在每个固定基上的投影来近似信号，接下来会介绍如何利用随机投影来近似信号。与采用固定基方式相比，这只需要较少的（随机的）基函数。这种方式可以与每个投影系数的 ℓ_1 惩罚项相结合，这就是**压缩感知**的思想。

信号 θ^* 的随机投影可以写成

$$y_i = \langle z_i, \theta^* \rangle = \sum_{j=1}^{p} z_{ij}\theta_j^* \tag{10.12}$$

其中，$z_i \in \mathbb{R}^p$ 是一个随机向量。使用随机投影来降维和近似的思想很早就有了，这（至少）可追溯到关于凸体（convex body）的度量嵌入（metric embedding）和球面部分的经典著作（详见本章文献注释）。这里首先介绍了一个经典的随机投影方法，即嵌入数据时保持点之间的距离，接下来讨论压缩感知，它将随机投影与 ℓ_1 松弛结合在一起。

10.3.1 Johnson-Lindenstrauss 近似

下面介绍一个随机投影的应用：使用随机投影近似有限的向量集，即表示某个数据集。这一方法通常称为 Johnson-Lindenstrauss 嵌入（Johnson-Lindenstrauss embedding），Johnson 和 Lindenstrauss 是使用它来研究更一般度量嵌入问题的先驱（详见本章文献注释）。假设在 $\mathbb{R}^p$ 中有 M 个数据点 $\{u_1, u_2, \ldots, u_M\}$，数据维度 p 较大，存储这个数据需要很大的空间。在这种情形下，需要设计一种降维映射：$F : \mathbb{R}^p \to \mathbb{R}^N, N \ll p$，映射后能保留数据集的“本质”特征，并且只保存投影的数据集 $\{F(u_1), F(u_2), \ldots, F(u_M)\}$。例如，由于许多算法需要在数据集上逐点求距离，这时人们可能会关注满足不等式

$$(1-\delta)\left\|u_i - u_{i'}\right\|_2^2 \leqslant \left\|F(u_i) - F(u_{i'})\right\|_2^2 \leqslant (1+\delta)\left\|u_i - u_{i'}\right\|_2^2,\ i \neq i' \tag{10.13}$$

的映射 F，其中 $\delta \in (0,1)$ 为容忍度。只要投影维度 N 足够大，这总有可能成立。但通常可以在 N 相对较小的情形下达到这样的目标。

Johnson 和 Lindenstrauss 的开创性工作表明，随机投影可作为这种近似距离保持嵌入的一种方法。其构造过程很简单。

(a) 得到一个随机矩阵 $\boldsymbol{Z} \in \mathbb{R}^{N \times p}$，其中每个 $Z_{ij} \sim N(0,1)$ 独立同分布，并定义线性映射 $F:\mathbb{R}^p \to \mathbb{R}^N$ 为

$$F(u) := \frac{1}{\sqrt{N}}\boldsymbol{Z}u \tag{10.14}$$

(b) 计算投影后的数据集 $\{F(u_1), F(u_2), \ldots, F(u_M)\}$。

这里有一个有趣的问题：对于给定的容忍度 $\delta \in (0,1)$ 和 M 个数据点，要选择多大的投影维度 N 才可有很高的概率让这种近似的距离保持属性（10.13）成立？习题 10.1 和习题 10.2 得证明：对于某个常数 c，只要 $N > \frac{c}{\delta^2}\log M$，就会以高概率使式（10.13）成立。这个条件与 M 的对数有关，这是一个非常宽松的限制条件。

下面来看一个具体的例子。假设要得到布尔向量 $u \in \{-1,1\}^p$ 的一个压缩表示，即使其成为一个 k 稀疏向量[①]。通过简单的计数就可以知道，这里有 $M = 2^k\binom{p}{k}$ 个这样的向量。注意，$\log M \leqslant k\log(\frac{e^2p}{k})$，则当投影维度为 $N > \dfrac{c}{\delta^2}k\log(\frac{e^2p}{k})$ 时，就可以保证所有 k 稀疏布尔向量之间具有 δ 精度的逐点距离。这就是一个压缩感知的例子，压缩感知会将随机投影与 ℓ_1 松弛结合在一起。

10.3.2 压缩感知

压缩感测分别由 Candes and Tao（2005）和 Donoho（2006）独立提出，它将随机投影与 ℓ_1 正则化结合在一起。这一开创性的工作产生了很多研究成果，并得到了广泛应用，包括医疗成像、单像素相机等。本节将简述压缩感知的基本思想。

提出压缩感知是为了解决基于正交基的标准信号压缩方法的浪费问题。10.2.2 节介绍的压缩会首先计算整个基系数向量 $\beta^* \in \mathbb{R}^p$，然后丢弃大部分系数来得到信号 θ^* 的 k 稀疏近似 $\hat{\theta}^k$。既然最终要丢弃大部分基系数，是不是真有必要计算所有这些系数呢？当然，若是事先知道由哪样的 k 系数子集可以得到稀疏近似，那么只计算基系数的这个子集即可。这种方式称为“最佳情况”技术。当然，这在实践中是不可能实现的，因为对于给定的信号，人们事先并不知道哪些系数最相关。

压缩感知可让人们用很少的计算代价来模仿“最佳情况”技术。它按如下方式将随机投影与 ℓ_1 最小化结合在一起：不用预先计算所有的基系数 $\beta^* = \boldsymbol{\Psi}^{\mathrm{T}}\theta^*$，而是计算 N 次随机投影数，即 $y_i = \langle \boldsymbol{z}_i, \theta^* \rangle$，$i = 1, 2, \ldots, N$。随机投影向量 $\boldsymbol{z}_i \in \mathbb{R}^p$ 可自由选择，下面简单讨论一些合理的选择。

这个问题描述为：有一个信号 θ^* 的 N 维随机投影向量 $\boldsymbol{y}$，还有一个用来计算随机投影的随机矩阵 $\boldsymbol{Z}$，其大小为 $N \times p$，称为设计矩阵（design matrix）或度量矩阵，它的第 i 行用 $\boldsymbol{z}_i$ 表示。观测到的向量 $\boldsymbol{y}$ 和设计矩阵 $\boldsymbol{Z}$ 之间通过未知信号 $\theta^* \in \mathbb{R}^p$ 联系在一起，即 $\boldsymbol{y} = \boldsymbol{Z}\theta^*$，这里的目标是（精确或近似地）恢复信号

① 当且仅当 $k \leqslant p$ 个元素为非零时，向量 $u \in \mathbb{R}^p$ 是 k 稀疏的。

$\theta^* \in \mathbb{R}^p$。图 10-5a 就是这种情形的示意图。

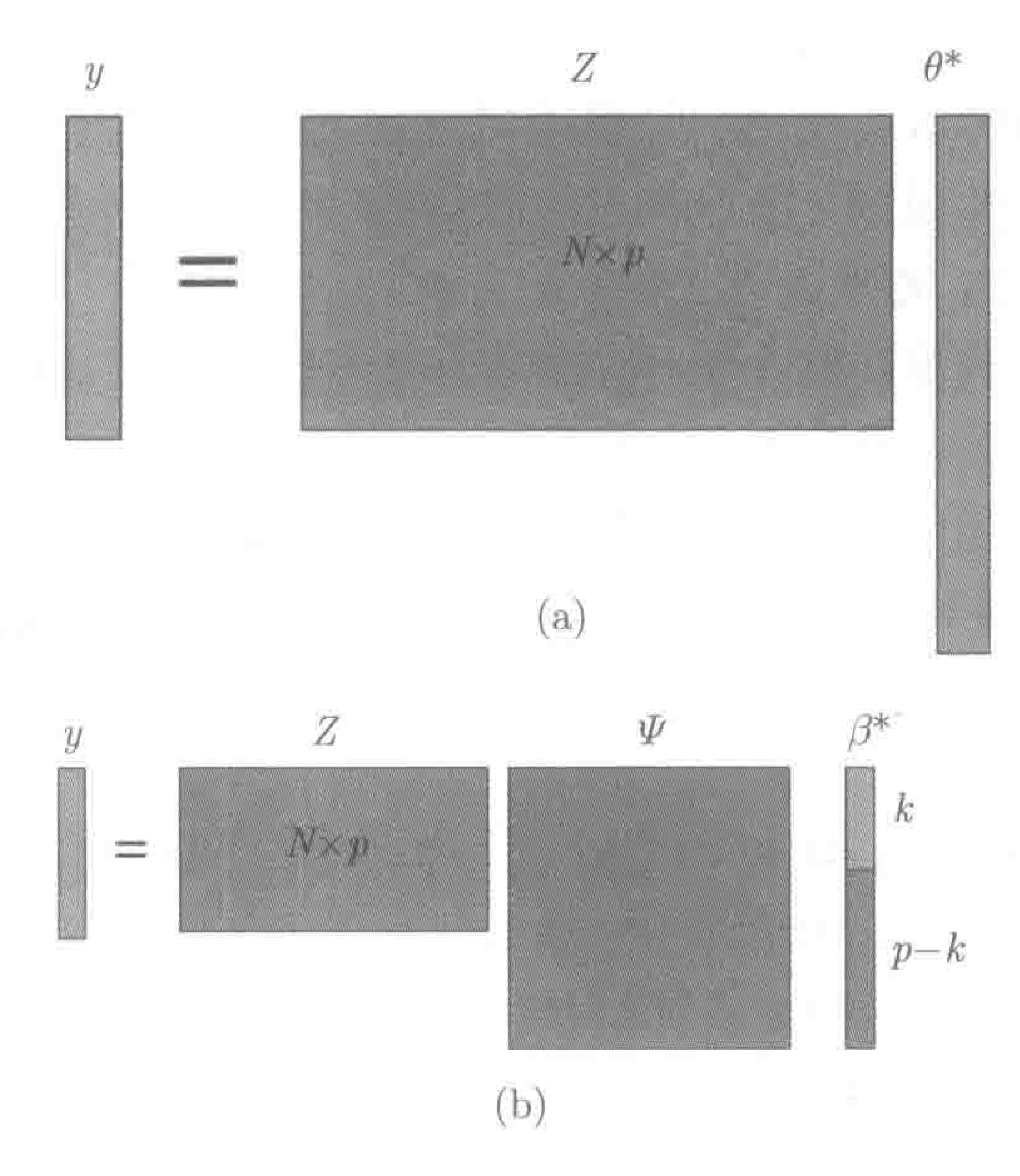


图 10-5 (a) 设欠定（under-determined）线性方程 $\boldsymbol{y} = \boldsymbol{Z}\theta^*$，度量矩阵 $\boldsymbol{Z}$ 的大小为 $N \times p$，它的第 i 行用 $\boldsymbol{z}_i$ 表示，由此可定义随机投影 $y_i = \langle \boldsymbol{z}_i, \theta^* \rangle$，信号 $\theta^* \in \mathbb{R}^p$ 在通常的基上不稀疏。(b) 线性方程的等价表示：假定基系数 $\beta^* = \boldsymbol{\Psi}^{\mathrm{T}}\theta^*$ 为 k 稀疏。这种变换定义了等价的线性方程 $\boldsymbol{y} = \tilde{\boldsymbol{Z}}\beta^*$，这个线性方程的解具有稀疏性

乍一看，这个问题似乎很简单，因为可以通过求解线性方程来得到 θ^*。但这里要提出一种比标准方法更简洁的方法。从本质上讲，投影数量（或样本数）N 要比维度 p 小得多。因此，线性方程 $\boldsymbol{y} = \boldsymbol{Z}\theta^*$ 有无穷多个解，其中的一些解可能与观测到的随机投影一致。

但如果附加一个信息：$\boldsymbol{\Psi}^{\mathrm{T}}\theta^*$ 是稀疏的，那么即使线性方程的解不唯一，仍可能精确恢复这个信号。理想情形是，通过求解基于 ℓ_0 的目标函数

$$\underset{\theta \in \mathbb{R}^p}{\operatorname{minimize}} \left\| \boldsymbol{\Psi}^{\mathrm{T}}\theta \right\|_0, \quad \text{使得} \boldsymbol{y} = \boldsymbol{Z}\theta \tag{10.15}$$

来得到这样的稀疏性。基于 ℓ_0 的问题是一个组合问题，通常计算起来很困难（NP 难）。因此要考虑该问题的 ℓ_1 松弛

$$\underset{\theta \in \mathbb{R}^p}{\operatorname{minimize}} \left\| \boldsymbol{\Psi}^{\mathrm{T}}\theta \right\|_1, \quad \text{使得} \boldsymbol{y} = \boldsymbol{Z}\theta \tag{10.16}$$

这个式子也可变换系数向量 $\beta \in \mathbb{R}^p$，写成另一种等价形式，即

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{minimize}} \|\beta\|_1, \quad \text{使得} \boldsymbol{y} = \tilde{\boldsymbol{Z}}\beta \tag{10.17}$$

其中，$\tilde{\boldsymbol{Z}} := \boldsymbol{Z}\boldsymbol{\Psi} \in \mathbb{R}^{N\times p}$，图 10-5b 给出了这种变换后的线性方程的示意图。

总之，压缩感知方法的实现过程如下：

(1) 对于给定的样本大小 N，计算随机投影 $y_i = \langle \boldsymbol{z}_i, \theta^* \rangle$，$i = 1, 2, \ldots, N$；

(2) 通过求解线性规划（10.16）来估算信号 θ^*，并由此得到 $\hat{\theta}$（这等价于通过求解线性规划（10.17）来得到 $\hat{\beta}$，并令 $\hat{\theta} = \boldsymbol{\Psi}\hat{\beta}$）。

需要明确的是，这里实际上已经介绍了一系列的算法，这些算法与随机投影向量 $\{z_i\}_{i=1}^N$ 的选择有关，或者说与设计矩阵 $\boldsymbol{Z}$ 有关。各种不同的设计矩阵 $\boldsymbol{Z}$ 被用于压缩感知的研究。也许最简单的选择，是让设计矩阵的元素独立同分布地服从 $z_{ij} \sim N(0, 1)$，这会得到标准的高斯随机矩阵。其他针对压缩感知的矩阵选择还包括元素具有独立同分布的随机伯努利矩阵，即该矩阵的元素 $z_{ij} \in \{-1, +1\}$；以及傅里叶矩阵的随机子矩阵。

当所使用的样本数 N 比信号维度 p 少很多时，压缩感知才会成功吗？10.4.2 节将讨论这方面的问题，它的一个充分条件是：变换后的设计矩阵 $\tilde{\boldsymbol{Z}}$ 的列要足够“不相关”，并对这些不相关有不同的度量。不相关的最简度量方式是让 $\tilde{\boldsymbol{Z}}$ 中的各列彼此做内积。一个更复杂的不相关概念与受限等距性（Restricted Isometry Property，RIP）有关，其基础是寻找 $\tilde{\boldsymbol{Z}}$ 的子矩阵条件，该子矩阵最多由 k 列所构成。一个重要的事实是，上面讨论的随机设计矩阵在样本数 N 相对较小的情况下，会以很高的概率满足 RIP 条件。例如，对于标准的高斯或伯努利情形，可以证明：样本数为 $N = \Omega(k \log \frac{p}{k})$ 时，RIP 成立的概率很高，其中 $k < p$ 表示基系数向量 β^* 的稀疏性。注意，任何方法，甚至已经知道 β^* 支持情况的“最佳情况”方法要做到精确恢复都需要至少 $N = K$ 个随机投影。因此，压缩感知得到的开销为 $\mathcal{O}(\log(p/k))$。

10.4 ℓ_0 恢复与 ℓ_1 恢复之间的等价性

到目前为止，本书已经讨论了信号处理中与 ℓ_1 范数正则化相关的许多应用，包括基于过完备基的稀疏近似（10.2.3 节）和压缩感知（10.3.2 节）。在这两种情况下，引入 ℓ_1 范数是为了解决 ℓ_0 范数难以计算的问题。至此，深层次的重要问题还没有涉及：什么时候求解 ℓ_1 松弛等价于求解原来的 ℓ_0 问题？

更准确地说，给定观测向量 $\boldsymbol{y} \in \mathbb{R}^p$ 和设计矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，这里要考虑两个问题

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \|\beta\|_0, \quad \text{使得} \boldsymbol{X}\beta = \boldsymbol{y} \tag{10.18}$$

和

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \|\beta\|_1, \quad \text{使得} \boldsymbol{X}\beta = \boldsymbol{y} \tag{10.19}$$

这包括过完备基上稀疏近似问题的特殊情形（10.2.3 节曾讨论过）。在这种情况下，观测向量 $\boldsymbol{y}$ 等于被近似信号 θ^*，而设计矩阵为 $\boldsymbol{X} = [\boldsymbol{\Phi} \quad \boldsymbol{\Psi}]$。这也包括了压缩感知，其中 $\boldsymbol{X}$ 是变换后的随机投影矩阵（即前面符号 $\tilde{\boldsymbol{Z}}$）。

10.4.1 受限零空间性质

假设基于 ℓ_0 问题（10.18）有唯一的最优解，即 $\beta^* \in \mathbb{R}^p$。这里需要明白，β^* 什么时候也是基于 ℓ_1 问题（10.19）的唯一最优解。在这种情况下，可认为基追踪 LP 等价于 ℓ_0 恢复。值得注意的是，对于设计矩阵 $\boldsymbol{X}$，这里存在一个非常简单的充分必要条件使这种等价成立。对于给定的子集 $S \subseteq \{1, 2, \ldots, p\}$，定义

$$\mathbb{C}(S) := \{\beta \in \mathbb{R}^p | \, \|\beta_{S^c}\|_1 \leqslant \|\beta_S\|_1\} \tag{10.20}$$

集合 $\mathbb{C}(S)$ 是一个凸锥，包含 S 上所有被支持的向量。简略地讲，它对应向量的锥体，这个锥体的大部分都分配给了 S。给定矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，其零空间为 $\text{null}(\boldsymbol{X}) = \{\beta \in \mathbb{R}^p | \boldsymbol{X}\beta = \boldsymbol{0}\}$。

定义 10.1：受限零空间性质　对于给定的子集 $S \subseteq \{1, 2, \ldots, p\}$，若

$$\text{null}(\boldsymbol{X}) \cap \mathbb{C}(S) = \{0\} \tag{10.21}$$

则设计矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ 在 S 上满足**受限零空间性质**，记为 RN(S)。也就是说，当 $\mathbb{C}(S)$ 与 RN(S) 相交为零向量时，该性质成立。下面的定理强调了这个性质的意义。

定理 10.1：$\boldsymbol{\ell_0}$ 与 $\boldsymbol{\ell_1}$ 的等价性　假设 $\beta^* \in \mathbb{R}^p$ 是 ℓ_0 问题（10.18）的唯一解，并且有支撑集 S。则通过基追踪松弛（10.19）所得的唯一解等于 β^*，当且仅当 $\boldsymbol{X}$ 满足 RN(S) 性质。

定理 10.1 的证明相对较短，将在 10.4.3 节给出。

因为子集 S 预先并不知道（这通常需要寻找），因此很自然要寻找一个矩阵，使其满足限制零空间性质的一致版本。比如，若 RN(S) 对大小最多为 k 的所有子集都成立，则 k 阶一致 RN 性质成立。在这种情况下，可保证 ℓ_1 松弛在任意被支撑的向量上成功。

10.4.2 受限零空间的充分条件

当然，为了使定理 10.1 在实际中有用，需要验证受限零空间的性质。其中的一些工作已经得到了在各种条件下一致 RN 性质。最简单且最早的条件称为成对不相关性（pairwise incoherence）

$$\nu(\boldsymbol{X}) := \max_{j,j'=1,2,\ldots,p} \frac{|\langle \boldsymbol{x}_j, \boldsymbol{x}_{j'} \rangle|}{\|\boldsymbol{x}_j\|_2 \, \|\boldsymbol{x}_{j'}\|_2} \tag{10.22}$$

对于中心化的 $\boldsymbol{x}_j$，这会得到成对不相关的最大绝对值。重新缩放 $\boldsymbol{X}$，使它的列具有单位范数，等价的表示为 $\nu(\boldsymbol{X}) = \max_{j\neq j'} |\langle \boldsymbol{x}_j, \boldsymbol{x}_{j'}\rangle|$，这表明成对不相关性会逐元素地度量 Gram 矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 与 p 维的单位矩阵的相似程度。

下面的结果表明，具有低的成对不相关性能充分保证基追踪 LP 的正确性。

命题 10.1：成对不相关性意味着 RN　假设对于某个整数 $k \in \{1,2,\ldots,p\}$，成对不相关性满足的界为 $\nu(\boldsymbol{X}) < \frac{1}{3k}$，则 $\boldsymbol{X}$ 满足 k 阶一致 RN 性质，因此，基追踪 LP 对最大为 k 的所有向量都能精确恢复，参见 10.4.3 节对这种命题的证明。

成对不相关性的一个迷人特性是容易计算，其时间复杂度为 $\mathcal{O}(Np^2)$。它的缺点是给出了非常保守的界，在应用中并不总是能得到 ℓ_1 松弛的实际性能。例如，过完备基问题（10.11）中的矩阵 $\boldsymbol{X} = [\boldsymbol{\Phi} \quad \boldsymbol{\Psi}]$ 可以通过数值计算来得到不相关性，即维度 $p = 128$ 的离散余弦基和哈尔基，如图 10-4 所示。命题 10.1 保证能精确恢复稀疏性为 $k = 1$ 的所有信号，而在实际应用中，ℓ_1 松弛需要更大的 k 值。

对于随机设计矩阵，比如压缩感知中的矩阵，可使用概率的方法来限制不一致性。例如，有一个随机矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，它的每个元素都是独立同分布的，并服从 $N(0,1/N)$。这里重新调整方差，使得 $\boldsymbol{X}$ 的每列的范数都等于 1。对于这样的矩阵，可以证明：当 (N,p) 趋于无穷时，有很高的概率使 $\nu(\boldsymbol{X}) \succsim \frac{\log p}{N}$（见习题 10.5）。与命题 10.1 相结合，可以得到结论：只要样本数 $N \succsim \frac{\log p}{N}$，就可以基于 ℓ_1 松弛（10.16）精确恢复稀疏性最多为 k 的信号。

事实上，对于随机设计矩阵和压缩感知，若使用**受限等距特性**，这个比例会提高。回想一下，不相关条件（10.22）是对设计矩阵 $\boldsymbol{X}$ 中各列之间正交性的一种度量。受限等距的概念要约束 $\boldsymbol{X}$ 的更大子矩阵，以便使其有接近正交的列。

定义 10.2：受限等距性　对容忍度 $\delta \in (0,1)$ 和整数 $k \in \{1,2,\ldots,p\}$，若满足

$$\left\|\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S - \boldsymbol{I}_{k\times k}\right\|_{\mathrm{op}} \leqslant \delta \tag{10.23}$$

则 $\mathrm{RIP}(k,\delta)$ 对基数（cardinality）为 k 的所有子集 $S \subset \{1,2,\ldots,p\}$ 成立。$\|\cdot\|_{\mathrm{op}}$ 表示算子范数，或矩阵的最大奇异值。由于 $\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S$ 是对称矩阵，则有等式

$$\left\|\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S - \boldsymbol{I}_{k\times k}\right\|_{\mathrm{op}} = \sup_{\|u\|_2=1}\left|u^{\mathrm{T}}(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S - \boldsymbol{I}_{k\times k})u\right| = \sup_{\|u\|_2=1}\left|\|\boldsymbol{X}_S u\|_2^2 - 1\right|$$

由此可知：当且仅当基数为 k 的所有子集 $S \subset \{1,2,\ldots,p\}$ 满足

$$\frac{\|\boldsymbol{X}_S u\|_2^2}{\|u\|_2^2} \in [1-\delta, 1+\delta]\,,\ u \in \mathbb{R}^k\backslash\{0\}$$

$\mathrm{RIP}(k,\delta)$ 成立。这就是受限等距的由来。

下面的结果表明：对于受限零空间，RIP 是一个充分条件。

命题 10.2: RIP 意味受限零空间 若 RIP$(2k,\delta)$（$\delta < 1/3$）成立，则 k 阶一致 RN 性质成立，因此在最多 k 个元素上，对被支持的所有向量，ℓ_1 松弛能精确恢复。

习题 10.8 将证明这个命题较为宽松的版本。通过观察可知，RIP$(2k,\delta)$ 条件会约束大量子矩阵，即总共有 $\binom{p}{2k}$ 个。另一方面，实际的 RIP 常数 δ 不依赖 k，这与成对不相关性的情形不一样。

从随机矩阵理论的结果可知：只要 $N \gtrsim k\log\frac{ep}{k}$，随机投影矩阵 $\boldsymbol{X}$ 的各种选择会以很高概率满足 RIP。对于其他集成矩阵，这种结论只适用于各元素独立同分布，且服从 $N\left(0,\frac{1}{N}\right)$ 的标准高斯随机矩阵 $\boldsymbol{X}$（习题 10.6 有详细的描述）。因此可以看出，基于 RIP 的方法为精确恢复比成对不相关性少得多的样本提供了一种保证。如上面所讨论的，这是在 $N \gtrsim k^2\log p$ 时。另一方面，RIP 的一个主要缺点（这与成对不相关性形成了鲜明对比）是，它在实际应用中非常难以验证，因为总共有 $\binom{p}{2k}$ 个子矩阵。

10.4.3 证明

下面对前面各节得出的结论进行证明。

1. 证明定理 10.1

首先假设 $\boldsymbol{X}$ 满足 RN(S) 性质。令 $\hat{\beta} \in \mathbb{R}^p$ 为基追踪 LP（10.19）的最优解，并定义误差向量 $\Delta := \hat{\beta} - \beta^*$。这里的目标是要证明 $\Delta = 0$，为此，需要证明 $\Delta \in \text{null}(\boldsymbol{X}) \cap \mathbb{C}(S)$。一方面，$\hat{\beta}$ 和 β^* 分别是 ℓ_0 问题和 ℓ_1 问题的最优解，因此有 $\boldsymbol{X}\beta^* = \boldsymbol{y} = \boldsymbol{X}\hat{\beta}$，于是有 $\boldsymbol{X}\Delta = 0$。另一方面，β^* 对基于 ℓ_1 的问题（10.19）也是可行的。由 $\hat{\beta}$ 是最优解可以得到 $\|\hat{\beta}\|_1 \leqslant \|\beta^*\|_1 = \|\beta^*_S\|_1$。令 $\hat{\beta} = \beta^* + \Delta$，则有

$$\|\beta^*_S\|_1 \geqslant \|\hat{\beta}\|_1 = \|\beta^*_S + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \geqslant \|\beta^*_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1$$

其中最后一个不等式成立用到了三角不等式。重排这些项可以得到 $\Delta \in \mathbb{C}(C)$，由假设可知 $\boldsymbol{X}$ 满足 RN(S) 条件，因此可以得到 $\Delta = 0$。

习题 10.4 会进行反向证明。

2. 证明命题 10.1

为了不失一般性（需要重新缩放），假定 $\|\boldsymbol{x}_j\|_2 = 1$，$j = 1,2,\ldots,p$。为了简化符号，设不相关性条件为 $\nu(\boldsymbol{X}) < \frac{\delta}{k}$（$\delta > 0$），证明的过程会验证 $\delta = 1/3$ 的充分性。

对于基数为 k 的任意子集 S，假设 $\beta \in \mathbb{C}(S)\backslash\{0\}$，可以证明 $\|\boldsymbol{X}\beta\|_2^2 > 0$，因此下界

$$\|\boldsymbol{X}\beta\|_2^2 \geqslant \|\boldsymbol{X}_S\beta_S\|_2^2 + 2\beta_S^{\mathrm{T}}\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_{S^c}\beta_{S^c} \tag{10.24}$$

有

$$
\begin{aligned}
2\left|\beta_S^{\mathrm{T}} \boldsymbol{X}_S^{\mathrm{T}} \boldsymbol{X}_{S^c} \beta_{S^c}\right| &\leqslant 2\left|\sum_{i \in S} \sum_{j \in S^c}\|\beta_i\|\|\beta_j\|\|\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle\|\right| \\
&\overset{\text{(i)}}{\leqslant} 2\|\beta_S\|_1\|\beta_{S^c}\|_1 \nu(\boldsymbol{X}) \\
&\overset{\text{(ii)}}{\leqslant} \frac{2\delta\|\beta_S\|_1^2}{k} \\
&\overset{\text{(iii)}}{\leqslant} 2\delta\|\beta_S\|_2^2
\end{aligned}
$$

不等式 (i) 的成立用到了成对不相关性的定义（10.22）。不等式 (ii) 成立利用了 $\nu(\boldsymbol{X})$ 假设中的有界性，以及 $\beta \in \mathbb{C}(S)$。不等式 (iii) 成立利用了这样的事实：由于 S 的基数最多为 k，根据 Cauchy–Schwarz 不等式有 $\|\beta_S\|_1^2 \leqslant k\|\beta_S\|_2^2$。因此，有

$$
\|\boldsymbol{X}\beta\|_2^2 \geqslant \|\boldsymbol{X}_S\beta_S\|_2^2 - 2\delta\|\beta_S\|_2^2 \tag{10.25}
$$

为了完成证明，还需要得到 $\|\boldsymbol{X}_s\beta_s\|_2^2$ 的下界。设 $\|\cdot\|_{\mathrm{op}}$ 表示矩阵的算子范数（最大奇异值），则有

$$
\|\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S - \boldsymbol{I}_{k\times k}\|_{\mathrm{op}} \leqslant \max_{i\in S}\sum_{j\in S\setminus\{i\}}|\langle x_i, x_j\rangle| \leqslant k\frac{\delta}{k} = \delta
$$

因此，$\|\boldsymbol{X}_s\beta_s\|_2^2 \geqslant (1-\delta)\|\beta_s\|_2^2$，结合式（10.25）可得到 $\|\boldsymbol{X}_s\beta_s\|_2^2 > (1-3\delta)\|\beta_s\|_2^2$，如声明的那样，$\delta = 1/3$ 足够了。

参考文献注释

对于用小波和其他多尺度基进行表示，有很多关于图像和其他信号稀疏性的文献（Field 1987、Ruderman 1994、Wainwright, Simoncelli and Willsky 2001、Simoncelli 2005）。有多篇论文（Donoho and Stark 1989、Chen et al. 1998、Donoho and Huo 2001、Elad and Bruckstein 2002、Feuer and Nemirovski 2003）讨论过基于过完备基的稀疏近似。图 10-2 给出了多尺度基的示意图，这被称为方向金字塔（steerable pyramid，见 Simoncelli and Freeman 1995）。随机投影在计算机科学和数值线性代数中广泛使用（Vempala 2004、Mahoney 2011、Pilanci and Wainwright 2014 等）。Johnson and Lindenstrauss（1984）采用随机投影证明了一个以他们名字命名的引理，该引理用来得到度量嵌入存在的条件。压缩感知分别由 Candès, Romberg amd Tao（2006）和 Donoho（2006）独立提出。Lustig, Donoho, Santos and Pauly（2008）讨论了压缩感知在医疗成像中的应用，而 Candès and Wakin（2008）讨论了压缩感知在信号处理中的各种应用。

Donoho and Huo（2001）、Feuer and Nemirovski（2003）、Cohen, Dahmen and DeVore（2009）分别讨论了受限零空间性质。多位研究人员（Donoho and Huo 2001、Elad and Bruckstein 2002、Feuer and Nemirovski 2003）针对受限零空间性质的充分条件，研究了过完备基下的成对不相关性和其他设计矩阵。Candès and Tao（2005）引入受限等距性作为受限零空间性质的较宽松的充分条件。对于随机矩阵与独立同分布的子高斯行，它遵从一致界（union bound）和随机矩阵理论标准（Davidson and Szarek 2001、Vershynin 2012）结果的一个组合，即样本大小 $N > ck\log\left(\frac{ep}{k}\right)$ 可确保以高概率满足 RIP。Baraniuk, Davenport, DeVore and Wakin（2008）指出了 RIP 和 Johnson–Lindenstrauss 引理之间的联系。习题 10.6 给出了相关的计算。Krahmer and Ward（2011）给出了部分矛盾，从而证明了受限等距可以用来建立 Johnson–Lindenstrauss 类型保证。

习　题

习题 10.1　卡方密度（Chi-squared concentration）。如果 $Y_1,\ldots,Y_N$ 是独立同分布的 $\mathcal{N}(0,1)$ 的随机变量，则随机变量 $Z := \sum_{i=1}^N Y_i^2$ 具有 N 个自由度的卡方分布（可简单记为 $Z \sim \mathcal{X}_N^2$）。

(a) 求证：对于所有 $\lambda \in [-\infty, 1/2]$，有

$$\mathbb{E}\left[\exp(\lambda(Z-d))\right] = \left[\frac{\mathrm{e}^{-\lambda}}{\sqrt{1-2\lambda}}\right]^d \tag{10.26}$$

(b) 使用式（10.26）来证明

$$\mathbb{P}\left[|Z-N| \geqslant tN\right] \leqslant 2\mathrm{e}^{-\frac{Nt^2}{32}}, \quad t \in (0, 1/2) \tag{10.27}$$

习题 10.2　Johnson–Lindenstrauss 近似。10.3.1 节讨论了距离保持嵌入问题。

(a) 求证：对每个索引 $j = 1, 2, \ldots, M$，变量 $N\|F(u_j)\|_2^2$ 是一个自由度为 N 的卡方分布。

(b) 对于任意的 $\delta \in (0,1)$，定义事件

$$\varepsilon(\delta) = \left\{\frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2}\right\} \in [1-\delta, 1+\delta], \quad i \neq j$$

利用习题 10.1 的结果和一致界证明，只要 $N > \frac{64}{\delta^2}\log M$，则有

$$\mathbb{P}\left[\varepsilon(\delta)\right] \geqslant 1 - 2\mathrm{e}^{-N}$$

习题 10.3　对给定的紧集 $\mathcal{A} \subset \mathbb{R}^p$，$\varepsilon$ 覆盖集是指：$\mathcal{A}$ 中元素的一个子集 $\{u_1,\ldots,u_M\}$，对于任意的 $u \in \mathcal{A}$，某个索引 $j \in \{1,\ldots,M\}$ 有 $\|u-u_j\|_2 \leqslant \varepsilon$。$\varepsilon$-packing 集是指：在 $\{1,\ldots,M'\}$ 中，$\mathcal{A}$ 中元素的一个子集 $\{v^1,\ldots,v^M\}$，有 $\left\|v^i-v^j\right\|_2 > \varepsilon$（$i \neq j$）。

(a) 求证：任意的最小 ε 覆盖集必定是 ε-packing 集；

(b) 求证：任意最大 2ε-packing 集必定是 ε 覆盖集；

(c) 思考欧几里得球 $\mathbb{B}(1) = \{u \in \mathbb{R}^p | \|u\|_2 = 1\}$，对每个 $\epsilon \in (0,1)$，证明存在最多有 $M = (1/\epsilon)^{cp}$ 个元素的 ε 覆盖集，其中 $c > 0$ 是某个给定的常量。**提示**：使用 (b) 的结果考虑 p 维欧几里得球的体积。

习题 10.4　这个习题要证明与定理 10.1 相反的结论，即对于所有 S 稀疏向量，如果 ℓ_1 松弛的唯一最优解等于 ℓ_0 的解，则集合 $\text{null}(\boldsymbol{X})\backslash\{0\}$ 与 $\mathbb{C}(S)$ 没有交集。

(a) 给定向量 $\beta^* \in \text{null}(\boldsymbol{X})\backslash\{0\}$，考虑基追踪问题

$$\underset{\beta\in\mathbb{R}^p}{\min\text{imize}} \|\beta\|_1\text{，使得} \boldsymbol{X}\beta = \boldsymbol{X}\begin{bmatrix}\beta_S^* \\ 0\end{bmatrix}$$

这个问题的唯一解 $\hat{\beta}$ 与下面向量 $\begin{bmatrix}0 \\ -\beta_{S^c}^*\end{bmatrix}$ 之间有什么关系？

(b) 通过 (a) 的结果证明 $\beta^* \notin \mathbb{C}(S)$。

习题 10.5　设 $\boldsymbol{X} \in \mathbb{R}^p$ 是随机矩阵，它的元素独立同分布于 $N\left(0, \frac{1}{N}\right)$，求证：对于给定的常量 c，只要 $N > ck^2\log p$，它就满足成对不相关条件 (10.22)。(**提示**：习题 10.1 的结果可能有用)。

习题 10.6　设 $\boldsymbol{X} \in \mathbb{R}^p$ 是随机矩阵，它的元素独立同分布于 $N\left(0, \frac{1}{N}\right)$。求证：只要 $N > ck\log(ep/k)$（$c > 0$ 充分大），受限等距性 RIP 会以很高的概率成立。

(a) 求证：存在常数 c_1 和 c_2 以最低为 $1 - c_1\mathrm{e}^{-c_2Nt^2}$（其中任何给定的子集 S 的基数为 $2k$，$t \in (0,1)$）的概率使如下不等式成立：

$$\left\|\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S - \boldsymbol{I}_{2k\times 2k}\right\|_{\text{op}} \leqslant t \tag{10.28}$$

(b) 令 $\mathbb{B}(1;S) = \{u \in \mathbb{R}^p | \|u\|_2 = 1, u_{S^c} = 0\}$ 表示欧几里得球与在 S 上支持的向量子空间之间的交集。令 $\{u_1, \ldots, u_M\}$ 为 $\mathbb{B}(1;S)$ 的 ε 覆盖集（习题 10.3 定义过），求证式 (10.28) 可由下面的界得到的概率至少为 $1 - c_3\mathrm{e}^{c_4N\varepsilon^2}$，其中 $t \in (0,1)$：

$$\max_{j=1,\ldots,M}\left|\|\boldsymbol{X}_Su_j\|_2^2 - 1\right| \leqslant \epsilon$$

(c) 用 (b) 和习题 10.3 的结论完成证明。

习题 10.7　ℓ_0 和 ℓ_1 球。本习题讨论 ℓ_0 球和 ℓ_1 球之间的联系，并证明与 ℓ_1 松弛的成功相关的包含性质。对于给定的整数 $r \in \{1,\ldots,p\}$，考虑两个子集

$$\mathbb{L}_0(r) := \mathbb{B}_2(1) \cap \mathbb{B}_0(r) = \{\theta \in \mathbb{R}^p | \, \|\theta\|_2 \leqslant 1, \ \|\theta\|_0 \leqslant r\}$$
$$\mathbb{L}_1(r) := \mathbb{B}_2(1) \cap \mathbb{B}_1(\sqrt{r}) = \{\theta \in \mathbb{R}^p | \, \|\theta\|_2 \leqslant 1, \ \|\theta\|_1 \leqslant r\}$$

令 $\overline{\text{conv}}$ 表示凸包的闭集（应用于集合）。

(a) 求证：$\overline{\text{conv}}\,(\mathbb{L}_0(r) \subseteq \mathbb{L}_1(r))$。

(b) 求证：$\mathbb{L}_1(r) \subseteq 2\overline{\text{conv}}(\mathbb{L}_0(r))$。

提示：(b) 问题更具挑战性，读者可能需要考虑两组支持函数。

习题 10.8　本习题要证明命题 10.2 的一个宽松版本。

(a) 对于基数为 k 的任何子集 S，$\mathbb{C}(S) \bigcap \mathbb{B}_2(1)$ 包含于集合 $\mathbb{L}_1(r)$，$r = 4k$。

(b) 求证：如果 $\text{RIP}(8k, \delta)$（$\delta < 1/4$）成立，则受限零空间性质成立。

提示：习题 10.7(b) 可能有用。

第 11 章　lasso 的理论结果

本章重点介绍与 lasso 相关的一些理论结果。这里会给出 ℓ_2 的非渐近界和 lasso 的预测误差，以及它在未知回归向量的支撑集上的性能。

11.1　引　　言

考虑标准线性回归模型的矩阵-向量形式

$$\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{w} \tag{11.1}$$

其中 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ 是模型（设计）矩阵，$\boldsymbol{w} \in \mathbb{R}^N$ 是噪声向量，$\beta^* \in \mathbb{R}^p$ 是未知系数向量。本章将给出带约束形式的 lasso

$$\underset{\|\beta\|_1 \leqslant R}{\text{minimize}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 \tag{11.2}$$

及拉格朗日形式的 lasso

$$\underset{\beta\in R^p}{\text{minimize}} \left\{ \frac{1}{2N} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda_N \|\beta\|_1 \right\} \tag{11.3}$$

之前曾讨论过，这两种二次规划问题之间可通过拉格朗日对偶建立起联系，其中 λ_N 可以解释为与约束 $\|\boldsymbol{\beta}\|_1 \leqslant R$ 有关的拉格朗日乘子。

11.1.1　损失函数类型

给定一个 lasso 估计 $\widehat{\beta} \in \mathbb{R}^p$，可通过多种方式来评估其质量。在一些情况下，$\widehat{\beta}$ 的预测性能是重点，所以需要计算预测损失函数

$$\mathcal{L}_{\text{pred}}(\hat{\beta}; \beta^*) = \frac{1}{N}\|\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta^*\|_2^2 \tag{11.4}$$

这对应于 $\widehat{\beta}$ 在给定样本 $\boldsymbol{X}$ 上的均方误差。在其他应用上，包括医学成像、遥感及压缩传感，未知向量 β^* 是重点，所以更适合考虑损失函数（即 ℓ_2 误差）

$$\mathcal{L}_2(\hat{\beta}; \beta^*) = \|\hat{\beta} - \beta^*\|_2^2 \tag{11.5}$$

这称为**参数估计损失函数**。最终的重点可能是变量选择或者支持恢复（support recovery），所以会采用损失函数

$$\mathcal{L}_{\text{vs}}(\hat{\beta}; \beta^*) = \begin{cases} 0, & \text{sgn}(\hat{\beta}_i) = \text{sgn}(\beta_i^*),\ i = 1, \ldots, p \\ 1, & \text{其他} \end{cases} \tag{11.6}$$

由此评估所估计的 $\widehat{\beta}$ 与 β^* 是否有同样的符号支撑。本章对这三种损失函数给出理论结果。

11.1.2 稀疏模型类型

对于一些方法（例如 lasso）的经典分析会固定协变量 p 的数目，然后让样本数目 N 趋向于无穷。尽管这类分析在一些情况下十分有用，但是在很多情况下，协变量 p 的数目可能和样本数目 N 的数量级相同，或者远远大于 N。比如，微阵列基因表达分析，其中会涉及 $p = 10\,000$ 个基因的，$N = 100$ 个样本，而社交网络的大量个体只有相对较少的关系。在这些情况下，基于 "固定 p，增大 N" 得出的理论结果是否能给实际使用的人提供有用的指导，值得怀疑。

因此，本章的目的是研究适用于高维情况的理论，这意味着维度 $p \gg N$。当然，如果模型缺少附加结构，则不会从拥有 p 维向量和有限样本的数据中得到有用的信息。事实上，当 $N < P$ 时，线性模型（11.1）无法确定，例如，不可能区分模型 $\beta^* = 0$ 和 $\beta^* = \Delta$，其中 $\Delta \in \mathbb{R}^p$ 有 p 个元素，是 $\boldsymbol{X}$ 的 N 维零空间。

因此，这有必要对未知回归向量 $\beta^* \in \mathbb{R}^p$ 加上约束，这里主要关注不同的稀疏约束。第一种情况是**硬稀疏**，其中假设 β^* 有至多 $k \leqslant p$ 项非零项。这种硬稀疏模型对于预测、ℓ_2 范数损失以及变量选择损失（11.6）都有意义。假设模型在 k 个系数上完全支持可能过于严格，所以这里也会考虑弱稀疏模型，这意味着 β^* 可以通过有较少非零项的向量来近似。这种方式可以正式定义为：对参数 $q \in [0,1]$ 和半径 $R_q > 0$，有集合

$$\mathbb{B}(R_q) = \{\beta \in \mathbb{R}^p | \sum_{j=1}^{p} |\beta_i|^q \leqslant R_q\} \tag{11.7}$$

这个集合称为半径[①] R_q 的 ℓ_q"球"，如图 11-1 中所示。对于 $q \in [0,1)$，这不是严格意义上的球，因为这是一个非凸集合。在 $q = 0$ 时，约束 $\beta^* \in \mathbb{B}(R_0)$ 等价于使 β^* 有至多 $k = R_0$ 个非零项。

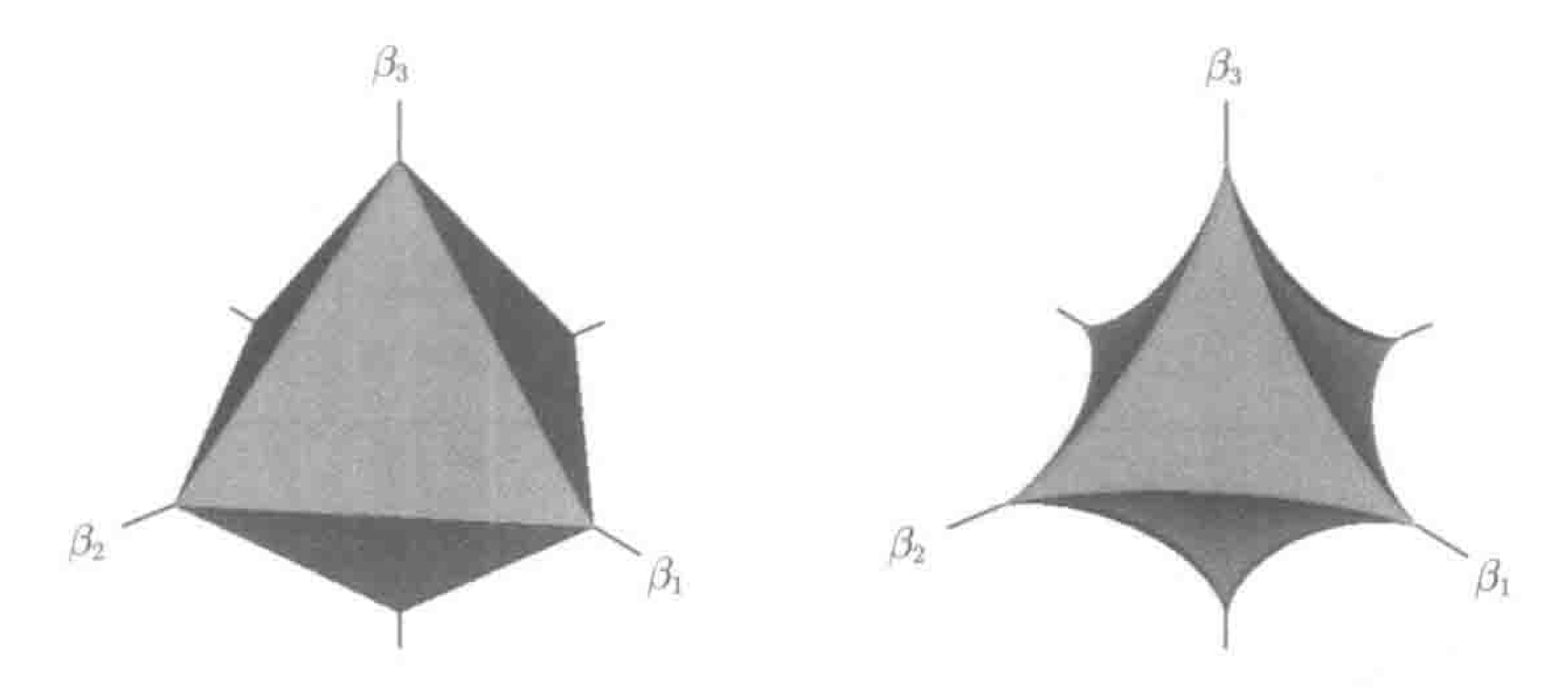


图 11-1 左图：对于 $q = 1$，集合 $\mathbb{B}(R_q)$ 对应 ℓ_1 球，这是一个凸集。右图：$q = 0.75$ 产生一个非凸集，沿着坐标轴正方向是尖的

① 严格来说，半径为 $R_q^{\frac{1}{q}}$，但是这里采用了简化的表示。

11.2 lasso ℓ_2 误差的界限

这里首先讨论 lasso 解 $\widehat{\beta}$ 和真实回归向量 β^* 之间的 ℓ_2 范数损失函数（11.5）的结果。讨论主要基于 β^* 为 k 稀疏的情形，这意味着它的非零元素的索引是基数为 $k = |S(\beta^*)|$ 的子集，$S(\boldsymbol{\beta}^*) \subset \{1, 2, \ldots, p\}$。习题会讨论扩展到弱稀疏系数向量的情形。

11.2.1 经典情形中的强凸性

下面先讨论关于模型矩阵 $\boldsymbol{X}$ 的一些条件，这些模型矩阵是确定 ℓ_2 误差界所需的。为了对这些条件给出一些直观的解释，需要先证明在经典情形（例如，固定 p，N 趋向于无穷）下 ℓ_2 的一致性。假设通过最小化定义在约束集上与数据有关的目标函数 $f_N(\beta)$，来估计某个参数向量 β^*。（例如，lasso 会最小化最小二乘损失函数 $f_N(\beta) = \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$，函数带一个 ℓ_1 约束。）随着样本数目 N 的增加，函数差异值 $\Delta f_N = |f_N(\widehat{\beta}) - f_N(\beta^*)|$ 收敛于零。这里需要解决的主要问题是：还需要什么附加条件来确保参数向量差值的 ℓ_2 范数 $\Delta\beta = \|\widehat{\beta} - \beta^*\|_2$ 也收敛于零？

为了理解所涉及的问题，这里假设对一些 N，目标函数 f_N 如图 11-2a 所示。由于目标函数在最小值 $\widehat{\beta}$ 附近相对“平”，因而可以看到，当参数差值 $\Delta\beta = \|\widehat{\beta} - \beta^*\|_2$ 相对较大时，函数差值 $\Delta f_N = |f_N(\widehat{\beta}) - f_N(\beta^*)|$ 很小。相反，图 11-2b 给出了一个更理想的情况，其中目标函数在最小值 $\widehat{\beta}$ 附近有很大的曲率。这种情况下，函数差值 $\Delta f_N = |f_N(\widehat{\beta}) - f_N(\beta^*)|$ 的界将会直接与 $\Delta\beta = \|\widehat{\beta} - \beta^*\|_2$ 的界有关。

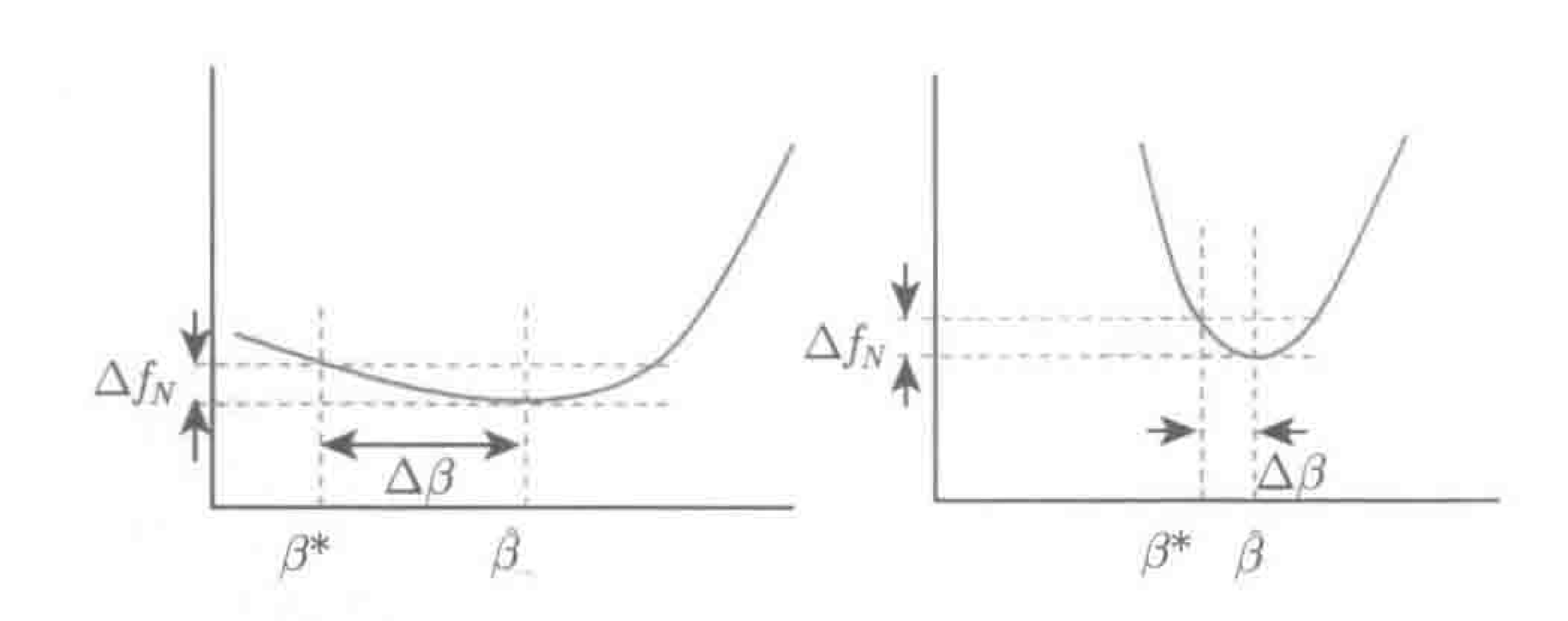


图 11-2 目标函数差值和参数差值之间的关系。左图：函数 f_N 在最优 $\widehat{\beta}$ 附近相对“平”，所以函数差值 $\Delta f_N = |f_N(\widehat{\beta}) - f_N(\beta^*)|$ 小并不意味着 $\Delta\beta = \|\widehat{\beta} - \beta^*\|_2$ 也小。右图：函数 f_N 在最优值附近极度弯曲，所以函数差值 Δf_N 小意味着参数差值小

如何将图 11-2 的这些情形正式表示出来呢？常用的方法是通过强凸性给出函数相应“弯曲”程度。具体而言，给定一个可微函数 $f : \mathbb{R}^p \to \mathbb{R}$，对于 $\theta \in \mathbb{R}^p$，$\gamma > 0$，若对于所有 $\theta' \in \mathbb{R}^p$，不等式成立，

$$f(\theta') - f(\theta) \geqslant \nabla f(\theta)^{\mathrm{T}}(\theta' - \theta) + \frac{\gamma}{2}\|\theta' - \theta\|_2^2 \tag{11.8}$$

则称该函数**强凸**。这个概念是普通凸函数的增强版，当 $\gamma = 0$ 时，强凸函数就是一个凸函数。当函数 f 二次连续可微，强凸性的另一特性与海森（Hessian）矩阵 $\nabla^2 f$ 有关。具体而言，当且仅当海森矩阵 $\nabla^2 f(\beta)$（对 β^* 邻域的所有向量 β）的最小特征值至少为 r 时，函数 f 是强凸的，其参数 r 在 $\beta^* \in \mathbb{R}^p$ 附近。如果 f 是一个参数模型下的负对数似然函数，则 $\nabla^2 f(\beta^*)$ 是观测到的 Fisher 信息矩阵，所以强凸性对应 Fisher 信息在所有方向上的一致下界（uniform lower bound）。

11.2.2 回归受限特征值

现在来介绍高维情况，其中参数 p 的数目可能大于 N。众所周知，最小二乘目标函数 $f_N(\beta) = \frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$ 总是凸的，它在什么附加条件下是强凸的呢？对于所有的 $\beta \in \mathbb{R}^p$，可以简单计算出 $\nabla^2 f(\beta) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/N$。因此，当且仅当 $p \times p$ 的半正定矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的特征值一致有界地远离零，最小二乘损失函数是强凸的。但是很容易看出，任意形如 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的矩阵的秩为 $\min\{N, p\}$，所以当 $N < p$ 时，该矩阵总是秩亏（rank-deficient）的，因此不强凸。图 11-3 说明了这一情况。

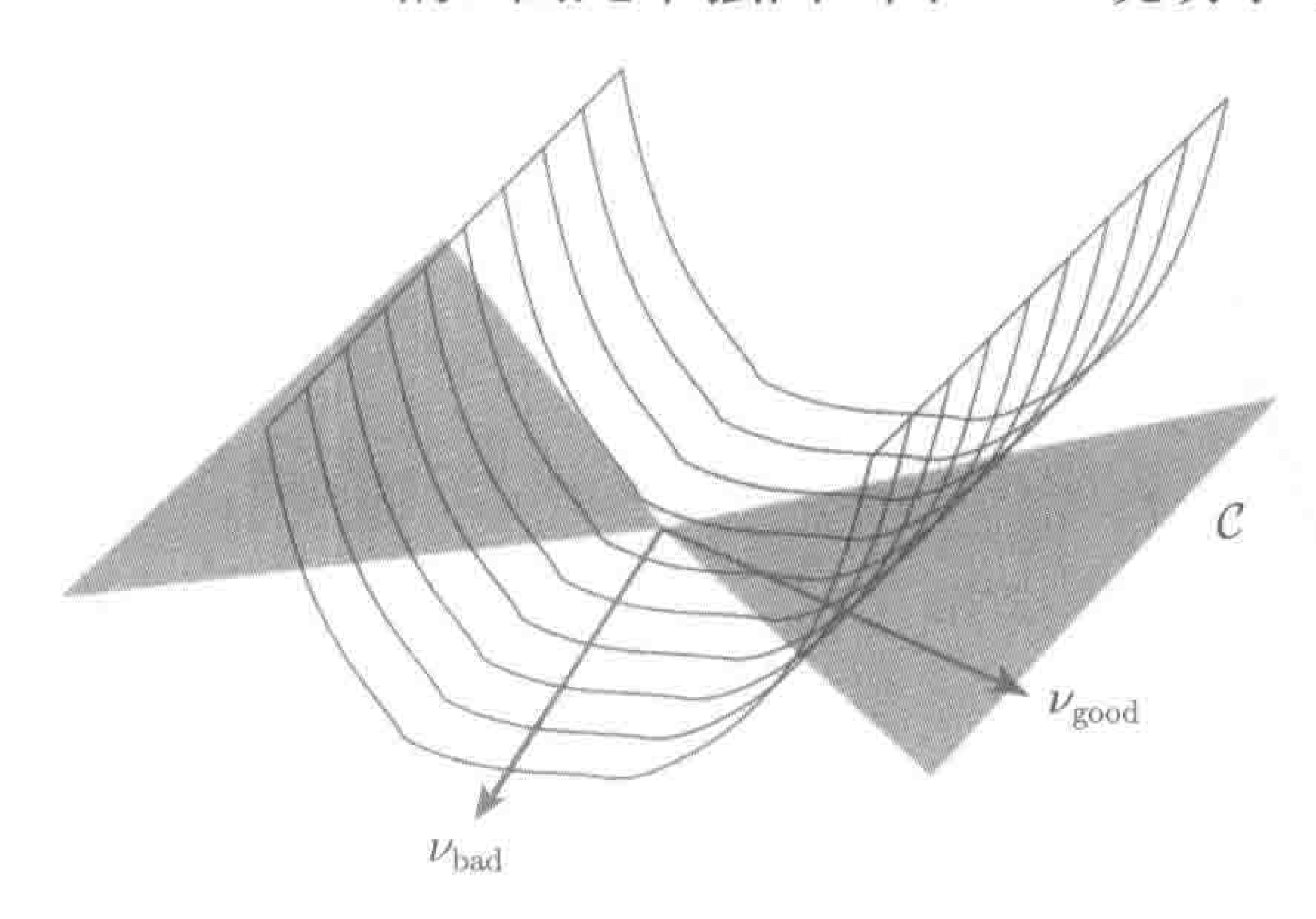


图 11-3　高维情况下（$p \gg N$）的凸损失函数不可能为强凸。相反，它在一些方向上是曲线（$\mathcal{V}_{\text{good}}$），而在另一些方向上是平的（$\mathcal{V}_{bad}$）。在引理 11.1 中，lasso 误差 $\widehat{\nu} = \widehat{\beta} - \beta^*$ 必须位于 $\mathbb{R}^p$ 的受限子集 $\mathbb{C}$ 中。因此，损失函数在空间的某些方向上弯曲是有必要的

这里需要放宽强凸的定义。下面的分析可说明，给可能的扰动向量 $\nu \in \mathbb{R}^p$ 的一些子集 $\mathcal{C} \in \mathbb{R}^p$ 加上一类强凸性条件即可。具体而言，函数 f 在 β^* 处关于 $\mathcal{C}$ 满足受限强凸性条件的充分条件是，有一个常数 $\gamma > 0$，对 β^* 邻域中所有的 $\beta \in \mathbb{R}^p$ 都满足

$$\frac{\nu^{\mathrm{T}}\nabla^2 f(\beta)\nu}{\|\nu\|_2^2} \geqslant \gamma，\quad \nu \in \mathcal{C} \tag{11.9}$$

对于线性回归，这等价于模型矩阵的**受限特征值**（restricted eigenvalue）的下界，即需要

$$\frac{\frac{1}{N}\nu \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\nu}{\|\nu\|_2^2} \geqslant \gamma, \quad \nu \in \mathcal{C} \tag{11.10}$$

这与约束集 $\mathcal{C}$ 有何相关? 假设参数向量 β^* 是稀疏的，即在子集 $S = S(\boldsymbol{\beta}^*)$ 上被支撑。定义 lasso 误差 $\widehat{\nu} = \widehat{\beta} - \beta^*$，设 $\widehat{\nu}_S \in \mathbb{R}^{|S|}$ 是以 S 为索引的子向量，并以类似的方式定义 $\widehat{\nu}_{S^c}$。要选择合适的 ℓ_1 球半径（等价于选择正则化参数 λ_N），lasso 误差应满足一个**锥形约束**，其形式为

$$\|\hat{\nu}_{S^c}\|_1 \leqslant \alpha \|\hat{\nu}_S\|_1 \tag{11.11}$$

其中，常数 $\alpha \geqslant 1$。这一事实对带约束的 lasso 显而易见。假设要求解带约束的 lasso（11.2），球半径为 $R = \|\beta^*\|_1$，因为 $\widehat{\beta}$ 是可用的，所以有

$$R = \|\beta_S^*\|_1 \geqslant \|\beta^* + \hat{\nu}\|_1 = \|\beta_S^* + \hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1 \geqslant \|\beta_S^*\|_1 - \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1$$

重新整理不等式，可以看到，α=1 时界成立。如果用一个“合适”的 λ_N 求解正则化版 lasso（11.3），其结果误差应满足约束

$$\|\hat{\nu}_{S^c}\|_1 \leqslant 3 \|\hat{\nu}_S\|_1 \tag{11.12}$$

（定理 11.1 的证明会得到这一事实。）因此，无论是约束还是正则形式，lasso 误差都限制在集合

$$\mathcal{C}(S;\alpha) : \{\nu \in \mathbb{R}^p |\, \|\nu_{S^c}\|_1 \leqslant \alpha \|\nu_S\|_1\} \tag{11.13}$$

中，其中参数 $\alpha \geqslant 1$，见图 11-3。

11.2.3 基本一致性结果

有了直观的概念，我们就可以开始介绍 lasso 误差 $\|\widehat{\beta} - \beta^*\|_2$ 的界了，这基于线性模型 $\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{w}$，其中 β^* 是 k 稀疏的，在子集 S 有支撑。

定理 11.1 假设模型矩阵 $\boldsymbol{X}$ 满足受限特征值界（11.10），在 $\mathcal{C}(S;3)$ 上参数 $\gamma > 0$，则

(a) 基于约束 lasso（11.2），$R = \|\beta^*\|_1$ 的任意估计量 $\widehat{\beta}$ 满足界

$$\|\hat{\beta} - \beta^*\|_2 \leqslant \frac{4}{\gamma}\sqrt{\frac{k}{N}}\|\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}}{\sqrt{N}}\|_\infty \tag{11.14a}$$

(b) 给定一个正则化参数 $\lambda_N \geqslant 2\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty/N > 0$，正则化 lasso（11.3）的任意估计量 $\widehat{\beta}$ 满足界

$$\|\hat{\beta} - \beta^*\|_2 \leqslant \frac{3}{\gamma}\sqrt{\frac{k}{N}}\sqrt{N}\lambda_N \tag{11.14b}$$

在证明这些结果之前，先讨论界 (11.14a) 和 (11.14b) 中的不同因子，然后用一些例子来说明。首先要注意，这些结果是确定的，可以运用到任意一组带观察噪声向量 $\boldsymbol{w}$ 的线性回归中。下面可得到一些特定统计模型的计算结果，这些结果依赖于噪声向量 $\boldsymbol{w}$ 和模型矩阵的假设。假设通过受限特征值常数 γ 和两个界中的项 $\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty$ 及 λ_N 会影响速率。基于之前强凸性作用的讨论，lasso 的 ℓ_2 误差反比于受限特征值常数 $\gamma>0$。第二项 $\sqrt{k/N}$ 也容易理解，因为这里基于 N 个样本来估计有 k 个未知项的回归向量。前面已经讨论，这两个界的最后项（涉及 $\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty$ 或者 λ_N）反映了观察噪声 $\boldsymbol{w}$ 和模型矩阵 $\boldsymbol{X}$ 之间的关系。

对于一些常用的线性回归模型，定理 11.1 的结论是有用处的。

例 11.1：经典线性高斯模型　下面从经典线性高斯模型开始介绍。设观察噪声 $\boldsymbol{w}\in\mathbb{R}^N$ 是高斯噪声，服从独立同分布 $N\left(0,\sigma^2\right)$。设计矩阵 $\boldsymbol{X}$ 是固定的，各列为 $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p\}$。对于任意给定的第 $j\in\{1,\ldots,p\}$ 列，通过简单的计算可以证明随机变量 $\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{w}/N$ 服从分布 $N(0,\frac{\sigma^2}{N}\cdot\frac{\|\boldsymbol{x}_j\|_2^2}{N})$。因此，如果设计矩阵 $\boldsymbol{X}$ 的列是正态的（意味着对所有的 $j=1,\ldots,p$，有 $\|\boldsymbol{x}_j\|_2/\sqrt{N}$），则这个变量由一个服从 $N\left(0,\frac{\sigma^2}{N}\right)$ 的随机变量决定，所以有高斯截尾界

$$\mathbb{P}\left[\frac{|\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{w}|}{N}\geqslant t\right]\leqslant 2\mathrm{e}^{-\frac{Nt^2}{2\sigma^2}}$$

因为 $\frac{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty}{N}$ 是 p 个这样变量上的最大值，所以由一致界可以得到

$$\mathbb{P}\left[\frac{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty}{N}\geqslant t\right]\leqslant 2\mathrm{e}^{-\frac{Nt^2}{2\sigma^2}+\log p}=2\mathrm{e}^{-\frac{1}{2}(\tau-2)\log p}$$

其中对 $\tau>2$，设 $t=\sigma\sqrt{\frac{\tau\log p}{N}}$ 就可让第二个等式成立。因此，可得出 lasso 误差的界

$$\left\|\hat{\beta}-\beta^*\right\|_2\leqslant\frac{c\sigma}{\gamma}\sqrt{\frac{\tau k\log p}{N}}\tag{11.15}$$

概率至少为 $1-2\mathrm{e}^{-\frac{1}{2}(\tau-2)\log p}$。这一计算也可以帮助选择对定理 11.1(b) 中拉格朗日 lasso 有效的正则化参数 λ_N。具体而言，在计算中设置 $\lambda_N=2\sigma\sqrt{\tau\frac{\log p}{N}}$ $(\tau>2)$，将是具有同样高概率的有效选择。

需要注意的是，式 (11.15) 直观上是合理的。事实上，如果支撑集 $S(\beta^*)$ 已知，则估计 β^* 大约需要 k 个参数，即所有 $i\in S(\beta^*)$ 的元素 β_i^*。即使支撑集已知，因为模型有 k 个自由参数，所以没法获得平方 ℓ_2 误差，该误差比 $\frac{k}{N}$ 衰退得更快。于是，除了对数因子，lasso 速率符合人们所能达到的最好情况，尽管子集 $S(\beta^*)$ 有**先验知识**。事实上，式 (11.15)（包括对数因子）是所谓的最小最大优化，这意味着它不可能由任何估计来提升。进一步的讨论见文献注释。

例 11.2: 压缩感知 在压缩感知领域（见第 10 章），设计矩阵 $\boldsymbol{X}$ 可由用户来选择，一个标准的方法就是通过服从独立同分布 $N(0,1)$ 的样本得到一个随机矩阵，模型中的噪声矩阵 $\boldsymbol{w}\in\mathbb{R}^N$ 是确定的，即有上界（$\|\boldsymbol{w}\|_\infty\leqslant\sigma$）在这些假设下，各个变量 $\frac{1}{N}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{w}$ 服从 0 均值高斯分布，方差最大为 $\frac{\sigma^2}{N}$。因此，通过前面例子中的相同参数，可得出结论：lasso 误差在这种情形下也以很高的概率服从界（11.15）。

更详细的论证可以推导出误差界（11.15）的加强版，即

$$\left\|\hat{\beta}-\beta^*\right\|_2\leqslant c\sigma\sqrt{\frac{k\log(\mathrm{e}p/k)}{N}} \tag{11.16}$$

其中 $\mathrm{e}\approx 2.71828$，$c$ 是一个全局常数。这个界建议样本数目 N 满足下界

$$N\geqslant k\log(\mathrm{e}p/k) \tag{11.17}$$

这会让使 lasso 有较小误差。

在 Donoho and Tanner（2009）工作的基础上，设比率 $\rho=k/N$，$\alpha=N/p$，这样，界（11.17）可以写成

$$\rho(1-\log(\rho\alpha))\leqslant 1 \tag{11.18}$$

为了研究预测的精度，这里生成了线性回归问题的随机组合，维度 $p=200$，样本数目 N 的范围是 10~200，其中特征 $x_{ij}\sim N(0,1)$ 独立生成。给定这个随机设计矩阵，从线性模型 $y_i=v\left\langle x_i,\beta^*\right\rangle+\sigma w_i$ 计算产生结果，其中 $w_i\sim N(0,1)$，$\sigma=4$。对于给定的稀疏水平 k，选择一个随机子集 S，数目为 k，对任意 $j\in S$，随机独立产生 $\beta_j^*\sim N(0,1)$。要在所有情况下，为每一个 N 和 k 选择预因子 ν，使得信噪比大约等于 10。接下来用正则化参数 $\lambda_N=2\sigma\sqrt{3\frac{\log\frac{\mathrm{e}p}{k}}{N}}$ 来求解拉格朗日 lasso。图 11-4 是欧几里得误差 $\|\widehat{\beta}-\beta^*\|_2$ 在 10 个样本上的中值热力图，加了边界条件（11.18）。可以看到，理论边界上的变化相当清楚，这意味模型更加稠密的时候需要更多的样本。

定理 11.1 的证明 对约束型 lasso 的界（11.14a）的证明十分简单，对正则型 lasso 界（11.14b）的证明比较麻烦。

约束型 lasso。在这种情况下，由于 β^* 是可行的，$\widehat{\beta}$ 是最优的，则有不等式 $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\beta}\|_2^2\leqslant\|\boldsymbol{y}-\boldsymbol{X}\beta^*\|_2^2$。定义误差向量 $\widehat{\nu}:=\widehat{\beta}-\beta^*$，代入 $\boldsymbol{y}=\boldsymbol{X}\beta^*+\boldsymbol{w}$，然后进行一些代数运算得到基本不等式（basic inequality）

$$\frac{\|\boldsymbol{X}\widehat{\nu}\|_2^2}{2N}\leqslant\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\widehat{\nu}}{N} \tag{11.19}$$

对右边运用 Hölder 不等式，可以得到上界 $\frac{1}{N}|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\widehat{\nu}|\leqslant\frac{1}{N}\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty\|\widehat{\nu}\|_1$。如第 10 章所示，不等式 $\|\widehat{\beta}\|_1\leqslant R=\|\beta^*\|_1$ 意味着 $\widehat{\nu}\in\mathcal{C}(S;1)$，因此有

$$\|\widehat{\nu}\|_1=\|\widehat{\nu}_S\|_1+\|\widehat{\nu}_{S^c}\|_1\leqslant 2\|\widehat{\nu}_S\|_1\leqslant 2\sqrt{k}\,\|\widehat{\nu}\|_2$$

另一方面，对不等式(11.19)的左边采用受限特征值条件(11.10)，会得到 $\frac{1}{N}\|\boldsymbol{X}\widehat{\nu}\|_2^2 \geqslant \gamma\|\widehat{\nu}\|_2^2$。将两部分合并得到界（11.14a）。

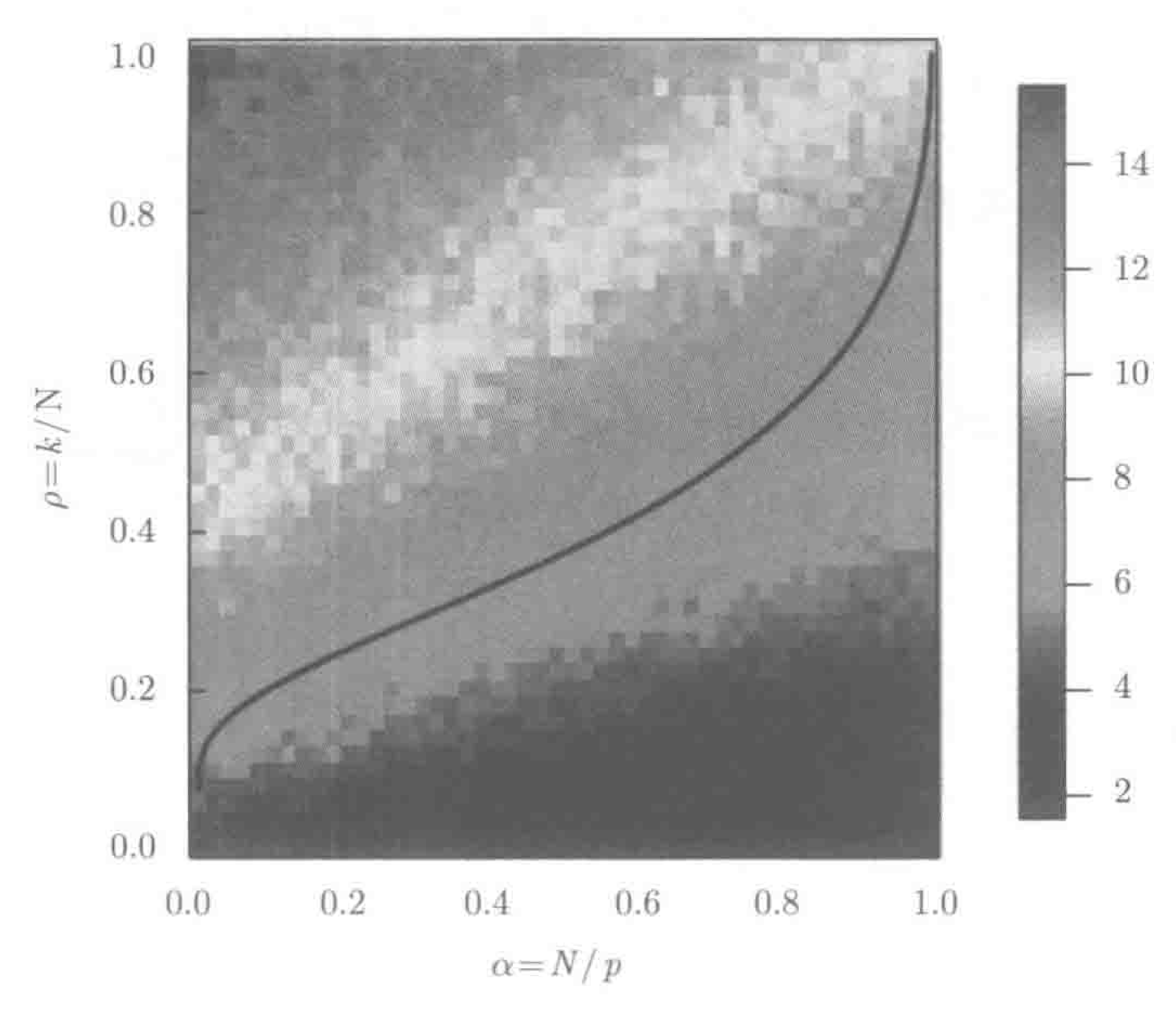


图 11-4 仿真实验：10 个样本上的误差 $\|\widehat{\beta}-\beta^*\|_2$ 中值，有界（11.18）（见彩插）

拉格朗日 lasso。 定义函数

$$G(\nu) := \frac{1}{2N}\|\boldsymbol{y}-\boldsymbol{X}(\beta^*+\nu)\|_2^2 + \lambda_N\|\beta^*+\nu\|_1 \tag{11.20}$$

注意，$\widehat{\nu} := \widehat{\beta}-\beta^*$ 通过构建来最小化 G，则有 $G(\widehat{\nu}) \leqslant G(0)$。进行代数运算得到改进型的基本不等式

$$\frac{\|\boldsymbol{X}\hat{\nu}\|_2^2}{2N} \leqslant \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\hat{\nu}}{N} + \lambda_N\left\{\|\beta^*\|_1 - \|\beta^*+\hat{\nu}\|_1\right\} \tag{11.21}$$

既然 $\beta^*_{S^c}=0$，则有 $\|\beta^*\|_1 = \|\beta^*_S\|_1$，以及

$$\|\beta^*+\hat{\nu}\|_1 = \|\beta^*_S+\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1 \geqslant \|\beta^*_S\|_1 - \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1$$

将这些关系代入不等式（11.21），就得到

$$\begin{aligned}\frac{\|\boldsymbol{X}\hat{\nu}\|_2^2}{2N} &\leqslant \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\hat{\nu}}{N} + \lambda_N\left\{\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1\right\} \\ &\leqslant \frac{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty}{N}\|\hat{\nu}\|_1 + \lambda_N\left\{\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1\right\}\end{aligned} \tag{11.22}$$

其中第二步使用了 ℓ_1 和 ℓ_∞ 范数的 Hölder 不等式。假设 $\frac{1}{N}\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty \leqslant \frac{\lambda_N}{2}$，则有

$$\frac{\|\boldsymbol{X}\hat{\nu}\|_2^2}{2N} \leqslant \frac{\lambda_N}{2}\left\{\|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^c}\|_1\right\} + \lambda_N\left\{\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1\right\} \leqslant \frac{3}{2}\sqrt{k}\lambda_N\|\hat{\nu}\|_2 \quad (11.23)$$

其中最后一步用到了 $\|\widehat{\nu}_S\|_1 \leqslant \sqrt{k}\,\|\widehat{\nu}\|_2$。

为了完成证明，需要下面的引理。

引理 11.1　假设 $\lambda_N \geqslant 2\|\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{N}}\|_\infty > 0$，则任意 lasso 解 $\widehat{\beta}$ 的误差 $\widehat{\nu} := \widehat{\beta} - \beta^*$ 属于锥集 $\mathcal{C}(S;3)$。

现在暂时假定该引理成立，以完成界（11.14b）的证明。引理 11.1 可以对 $\widehat{\nu}$ 采用 γ-RE 条件（11.10），这有 $\frac{1}{N}\|\boldsymbol{X}\widehat{\nu}\|_2^2 \geqslant \gamma\|\widehat{\nu}\|_2^2$。结合这个下界和之前的不等式（11.23），则有

$$\frac{\gamma}{2}\|\widehat{\nu}\|_2^2 \leqslant \frac{3}{2}\lambda_N\sqrt{k}\,\|\widehat{\nu}\|_2$$

然后重新整理得到界（11.14b）。

接下来证明引理 11.1。因为 $\frac{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty}{N} \leqslant \frac{\lambda_N}{2}$，所以不等式（11.22）意味着

$$0 \leqslant \frac{\lambda_N}{2}\|\widehat{\nu}\|_1 + \lambda_N\left\{\|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1\right\}$$

重新整理，分离出 $\lambda_N > 0$，即得到 $\|\widehat{\nu}_{S^c}\|_1 \leqslant 3\,\|\widehat{\nu}_S\|_1$。

扩展阅读。如前所述，定理 11.1 可以运用于回归模型，其中 β^* 至多含有 k 个非零项，这种假设称为硬稀疏。但是，一种相似的方法也可以用于弱稀疏模型，即 β^* 属于之前定义在式（11.7）中的 ℓ_q 球 $\mathbb{B}_q(R_q)$。在一组相似的假设下，可以证明 lasso 误差以很高的概率满足界

$$\left\|\hat{\beta} - \beta^*\right\|_2^2 \leqslant cR_q\left(\frac{\sigma^2\log p}{N}\right)^{1-q/2} \quad (11.24)$$

习题 11.3 会完成这一部分的推导工作。当 $q = 0$ 时，$\boldsymbol{\beta}^*$ 属于 ℓ_0 球的假设等价于硬稀疏的假设（半径 $R_0 = k$），所以这个比率（11.24）等价于之前从定理 11.1 中推导出的结果（11.16）。另外要注意，随着弱稀疏参数 q 从 0 向 1 增加，速率降低，这意味着真实回归向量 β^* 上加了弱条件。速率（11.24）是 ℓ_q 球上的极大极小优化，即没有其他的估计能达到如此小的 ℓ_2 误差，详细讨论见文献注释。

11.3　预测误差的界

目前为止，我们研究了 lasso 在计算真实回归向量上的表现，这均是通过欧几里得误差 $\|\widehat{\beta} - \beta^*\|_2$ 进行评价的。在其他情况下，这足以获得估计 $\widehat{\beta}$，该估计有相

对较低的（样本内）预测误差 $\mathcal{L}_{\text{pred}}(\widehat{\beta},\beta^*)=\frac{1}{N}\|\boldsymbol{X}(\widehat{\beta}-\beta^*)\|_2^2$。本节会就这种形式的损失函数讨论一些理论保证。具体而言，这里主要讨论拉格朗日 lasso（11.3），其他形式的 lasso 也可以得出相似的结果。

定理 11.2 考虑拉格朗日 lasso，正则化参数 $\lambda_N\geqslant\frac{2}{N}\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty$。

(a) 如果 $\|\beta^*\|_1\leqslant R_1$，则任意最优解 $\widehat{\beta}$ 满足

$$\frac{\|\boldsymbol{X}(\hat{\beta}-\beta^*)\|_2^2}{N}\leqslant 12R_1\lambda_N \tag{11.25a}$$

(b) 如果 β^* 由子集 S 支撑，设计矩阵 $\boldsymbol{X}$ 在 $\mathcal{C}(S;3)$ 上满足 γ-RE 条件（11.10），则任意最优解 $\widehat{\beta}$ 满足

$$\frac{\|\boldsymbol{X}(\hat{\beta}-\beta^*)\|_2^2}{N}\leqslant\frac{144}{\gamma}|S|\,\lambda_N^2 \tag{11.25b}$$

正如之前所讨论的，在多种统计模型中，对于定理 11.2，$\lambda_N=c\sigma\sqrt{\dfrac{\log p}{N}}$ 是高概率成立的，所以两个界的形式为

$$\frac{\|\boldsymbol{X}(\hat{\beta}-\beta^*)\|_2^2}{N}\leqslant c_1\sigma R_1\sqrt{\frac{\log p}{N}} \tag{11.26a}$$

$$\frac{\|\boldsymbol{X}(\hat{\beta}-\beta^*)\|_2^2}{N}\leqslant c_2\frac{\sigma^2}{\gamma}\frac{|S|\log p}{N} \tag{11.26b}$$

基于半径 R_1 的 ℓ_1 球的界（11.26a），对 lasso 而言是“慢速率”，因为平方预测误差以 $1/\sqrt{N}$ 衰减。而界（11.26b）是“快速率”，因为它以 $1/N$ 衰减。注意，后者基于一个强得多的假设：硬稀疏条件 β^* 由一个小的子集 S 支撑，而且在设计矩阵 $\boldsymbol{X}$ 上需满足 γ-RE 条件。原则上，预测性能并不需要 RE 条件，所以有人怀疑这个要求是出于证明方法的需要。文献注释会阐明，这种依赖性是任何多项式时间方法所无法避免的。

定理 11.2 的证明 这两个声明的证明相对来说较为简单。

界（11.25a）的证明 从改进版的基本不等式（11.21）开始，有

$$\begin{aligned}
0&\leqslant\frac{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty}{N}\|\hat{\nu}\|_1+\lambda_N\left\{\|\beta^*\|_1-\|\beta^*+\hat{\nu}\|_1\right\}\\
&\leqslant\left\{\frac{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty}{N}-\lambda_N\right\}\|\hat{\nu}\|_1+2\lambda_N\|\beta^*\|_1\\
&\overset{\text{(i)}}{\leqslant}\frac{1}{2}\lambda_N\left\{-\|\hat{\nu}\|_1+4\|\beta^*\|_1\right\}
\end{aligned}$$

其中，步骤（i）用到了 $\frac{1}{N}\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty\leqslant\lambda_N$。将两者合并，就可得出 $\|\hat{\nu}\|_1\leqslant 4\|\beta^*\|_1\leqslant$

$4R_1$。再一次回到改进的基础不等式（11.21），则有

$$\frac{\|\boldsymbol{X}\hat{\nu}\|_2^2}{2N} \leqslant \left\{\frac{\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|_\infty}{N} + \lambda_N\right\}\|\hat{\nu}\|_1 \leqslant 6\lambda_N R_1$$

这就证明了式（11.25a）。

界（11.25b）式的证明 给定选择好的 λ_N，使不等式（11.22）成立，由此得出

$$\frac{\|\boldsymbol{X}\hat{\nu}\|_2^2}{N} \leqslant 2\left\{\left\|\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}}{N}\right\|_\infty + \lambda_N\right\}\|\hat{\nu}\|_1 \leqslant 12\lambda_N\sqrt{k}\,\|\hat{\nu}\|_2$$

根据引理 11.1，误差向量 $\hat{\nu}$ 属于锥集 $\mathcal{C}(S;3)$，所以 γ-RE 条件保证了 $\|\hat{\nu}\|_2^2 \leqslant \frac{1}{N\gamma}\|\boldsymbol{X}^{\mathrm{T}}\hat{\nu}\|_2^2$。将这两部分合并得到式（11.25b）。

11.4 线性回归中的支持恢复

目前为止，本章主要讨论了 ℓ_2 误差或 lasso 解下的预测误差的界。在其他情况下，这里只关注了某个具体问题，如 lasso 估计 $\widehat{\beta}$ 是否如真实回归向量 β^* 一样在相同的位置有非零项。具体而言，假设真实回归向量 β^* 是 k 稀疏的，即它由基数为 $k=|S|$ 的子集 $S=S(\beta^*)$ 支持。在这种情形下，目标自然就是正确地找出相关变量的子集 S。对于 lasso 可以提出以下问题：给定最优 lasso 解 $\widehat{\beta}$，什么时候它的支撑集（用 $\widehat{S}=S(\widehat{\beta})$ 表示）正好等于其真实的支撑集 S？这个性质叫作**变量选择一致性**（**或稀疏性**）。

注意，即使 $\widehat{\beta}$ 和 β^* 有不同的支撑集，只要 $\widehat{\beta}$ 中的非零项相对所有 β^* 的元素“适当大”，且在 β^* 为零的位置不是“特别大”，则 ℓ_2 误差 $\|\widehat{\beta}-\beta^*\|_2$ 可能特别小。同样，即使 $\widehat{\beta}$ 和 β^* 有非常不同的支撑集，预测误差 $\|\boldsymbol{X}(\widehat{\beta}-\beta^*)\|_2/\sqrt{N}$ 也可能很小。另一方面，给定一个能正确恢复 β^* 的支撑的估计 $\widehat{\beta}$，就能很好地（在 ℓ_2 范数和预测半范数情况下）估计 β^*，即通过在这个子集上做普通最小二乘回归来很好地估计 β^*。

11.4.1 lasso 的变量选择一致性

首先，在确定设计矩阵 $\boldsymbol{X}$ 的情况下来讨论变量选择问题。变量选择需要的条件与受限特征值条件（11.10）相关，但又有些不同。特别是，要假设一个条件，即**互不相关性**（mutual incoherence）或者**无代表性**（irrepresentability）：必须存在某个 $\gamma>0$，使得

$$\max_{j\in S^c}\left\|(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{x}_j\right\|_1 \leqslant 1-\gamma \tag{11.27}$$

要理解这个条件，需要先注意一点：子矩阵 $\boldsymbol{X}_S \in \mathbb{R}^{N\times k}$ 对应的是支撑集中的协变量的子集。对互补集 S^c 中的任意 j，k 向量 $(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{x}_j$ 是 $\boldsymbol{x}_j$ 在 $\boldsymbol{X}_S$ 上的回归系数，这个向量衡量的是列 $\boldsymbol{x}_j$ 与子矩阵 $\boldsymbol{X}_S$ 的列对齐程度。在最理想的情况下，列 $\{\boldsymbol{x}_j, j\in S^c\}$ 与 $\boldsymbol{X}_S$ 的列正交，并保证 $\gamma=1$。当然，在高维情况下（$p\gg N$），不可能有完全正交性，但是仍要有“近似正交性”。

除了不相关性假设，这里也假设设计矩阵有归一化的列

$$\max_{j=1,\ldots,p}\|\boldsymbol{x}_j\|_2/\sqrt{N}\leqslant K_{\mathrm{clm}} \tag{11.28}$$

例如，可以取 $\|\boldsymbol{x}_j\|_2=\sqrt{N}$，$K_{\mathrm{clm}}=1$，再假设子矩阵 $\boldsymbol{X}_S$ 是好的，也就是说

$$\lambda_{\min}(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S/N)\geqslant C_{\min} \tag{11.29}$$

注意，如果这个条件不满足，那么 $\boldsymbol{X}_S$ 的列将线性相关，即使在“最佳情况”（即支撑集 S 已知），也不可能估计出 β^*。

以下结果适用于正则化 lasso（11.3），当用于线性观测模型（11.1）时，真实参数 β^* 的支撑数目为 k。

定理 11.3 假设设计矩阵 X 满足互不相关性条件（11.27）（参数 $\gamma>0$），列归一化条件（11.28）和特征值条件（11.29）都成立。对于一个噪声向量 $\boldsymbol{w}\in\mathbb{R}^N$，其服从独立同分布 $N(0,\sigma^2)$，考虑正则化 lasso（11.3），有

$$\lambda_N\geqslant\frac{8K_{\mathrm{clm}}\sigma}{\gamma}\sqrt{\frac{\log p}{N}} \tag{11.30}$$

则 lasso 以大于 $1-c_1\mathrm{e}^{-c_2N\lambda_N^2}$ 的概率拥有下列特性。

(a) **唯一性**：最优解 $\widehat{\beta}$ 是唯一的。

(b) **没有假包含性**（false inclusion）：唯一最优解的支撑集 $S(\widehat{\beta})$ 包含于真实支撑集 $S(\beta^*)$。

(c) **$\boldsymbol{\ell_\infty}$ 界**：误差 $\widehat{\beta}-\beta^*$ 满足 ℓ_∞ 界

$$\|\hat{\beta}_S-\beta_S^*\|_\infty\leqslant\underbrace{\lambda_N\left[\frac{4\sigma}{\sqrt{C_{\min}}}+\|(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S/N)^{-1}\|_\infty\right]}_{B(\lambda_N,\sigma;\boldsymbol{X})} \tag{11.31}$$

(d) **没有假排除性**（false exclusion）：lasso 的解包含所有的 $j\in S(\beta^*)$，$|\beta_j^*|>B(\lambda_N,\sigma;\boldsymbol{X})$，因此只要 $\min\limits_{j\in S}|\beta_j^*|>B(\lambda_N,\sigma;\boldsymbol{X})$，lasso 解就具有变量选择一致性。

在证明之前，先大致解释一下这些结论。首先，(a) 中的唯一性在高维情况下很重要。因为，虽然 lasso 目标函数是凸的（如面所讨论的），但是当 $p>N$ 时，它不可能是严格凸的。唯一性很重要，因为它可以让我们清楚地讨论由 lasso 估计得到的 $\widehat{\beta}$ 的支撑集。(b) 保证了 lasso 不会错误地包含不在 β^* 的支撑集中的变量，即 $\widehat{\beta}_{S^c}=0$，而 (c) 保证 $\widehat{\beta}_S$ 在 ℓ_∞ 范数上一致接近于 β_S^*。(d) 是一致范数界的结

果，只要 $|\beta_j^*|$（$j \in S$）的最小值不是太小，则 lasso 在完全意义上有变量选择一致性。

一些数值研究

为了进一步研究实际应用中这些结果的影响，这里可以进行一些小的仿真实验。首先研究互不相关性条件（11.27）的影响。给定样本数目 $N = 1000$，对一组维数 p，生成 p 个独立同分布标准高斯变量，$f = k/p$ 在支撑集 S 中。相关性 ρ 在区间 [0, 0.6] 内，对任意 $j \in S$，随机选择一个预测子 $\ell \in S^c$ 并置 $\boldsymbol{x}_\ell \leftarrow \boldsymbol{x}_\ell + c \cdot \boldsymbol{x}_j$，$c$ 可选，所以 $\operatorname{corr}(\boldsymbol{x}_j, \boldsymbol{x}_\ell) = \rho$。图 11-5 为 5 个样本集上 1-$\gamma$ 的平均值，即互不相关

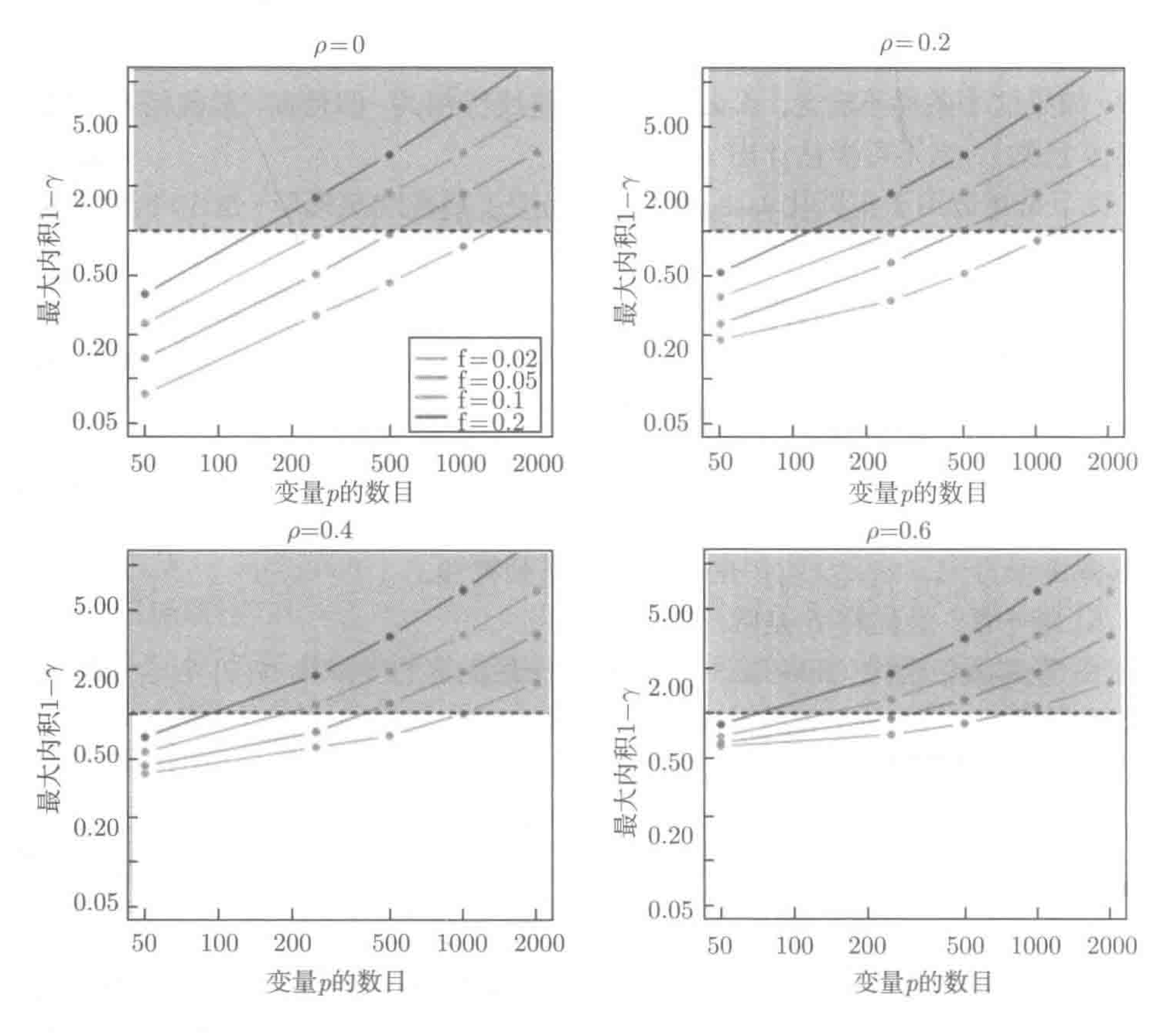


图 11-5　实际应用中的互不相关性条件。各图显示了模拟高斯数据下式（11.27）中 $1-\gamma$ 的值。比 1 小的值为好值，且越小越好。样本数 $N = 1000$ 是固定的，预测子数目 p 沿着水平轴而变化。真实非零系数的 $f = k/p$（稀疏水平）在各个图中变化，最终，真实预测子及其空预测子组（如文中描述）间的相关性在四幅图中都有变化。水平虚线画在 $1-\gamma = 1$ 处，在其下满足互不相关性条件。各个点是 5 个模拟上 $1-\gamma$ 的均值。均值的标准差较小，平均大约为 0.03

性条件（11.27）的值。以 $\rho=0$ 为例，$p\leqslant 1000$ 意味着进入“好”区域，这时稀疏性 $f\leqslant 2\%$，或者 $p\leqslant 500$ 有 $f\leqslant 5\%$ 的稀疏性。但是，p 的最大值及稀疏性水平 f 随着相关性 ρ 的增大而变小。

下面采用一个小的仿真实验来检测 lasso 回归下的假发现率（false discovery）和假排除率。设 $N=1000$，$p=500$，在 S 中有 $k=15$ 个预测子的系数不等于零。数据矩阵 $\boldsymbol{X}_S$ 和 $\boldsymbol{X}_{S^c}$ 按照上述条件生成，其中相关系数 ρ 有不同值。接下来通过 $\boldsymbol{y}=\boldsymbol{X}_S\beta_S+\boldsymbol{w}$ 得到输出 $\boldsymbol{y}$，$\boldsymbol{w}$ 的元素服从 N (0,1) 独立同分布。

这里对 β_S 中的非零回归系数尝试两种不同的值：所有都为 0.25 或者所有都为 0.15，随机选择符号。这样产生了 15 个真实预测子，其“有效大小”（绝对标准回归系数）分别为 7.9 和 4.7。最终，λ_N 在每个循环中以“最优”方式选择：采用产生正确非零系数数目（15）的值。

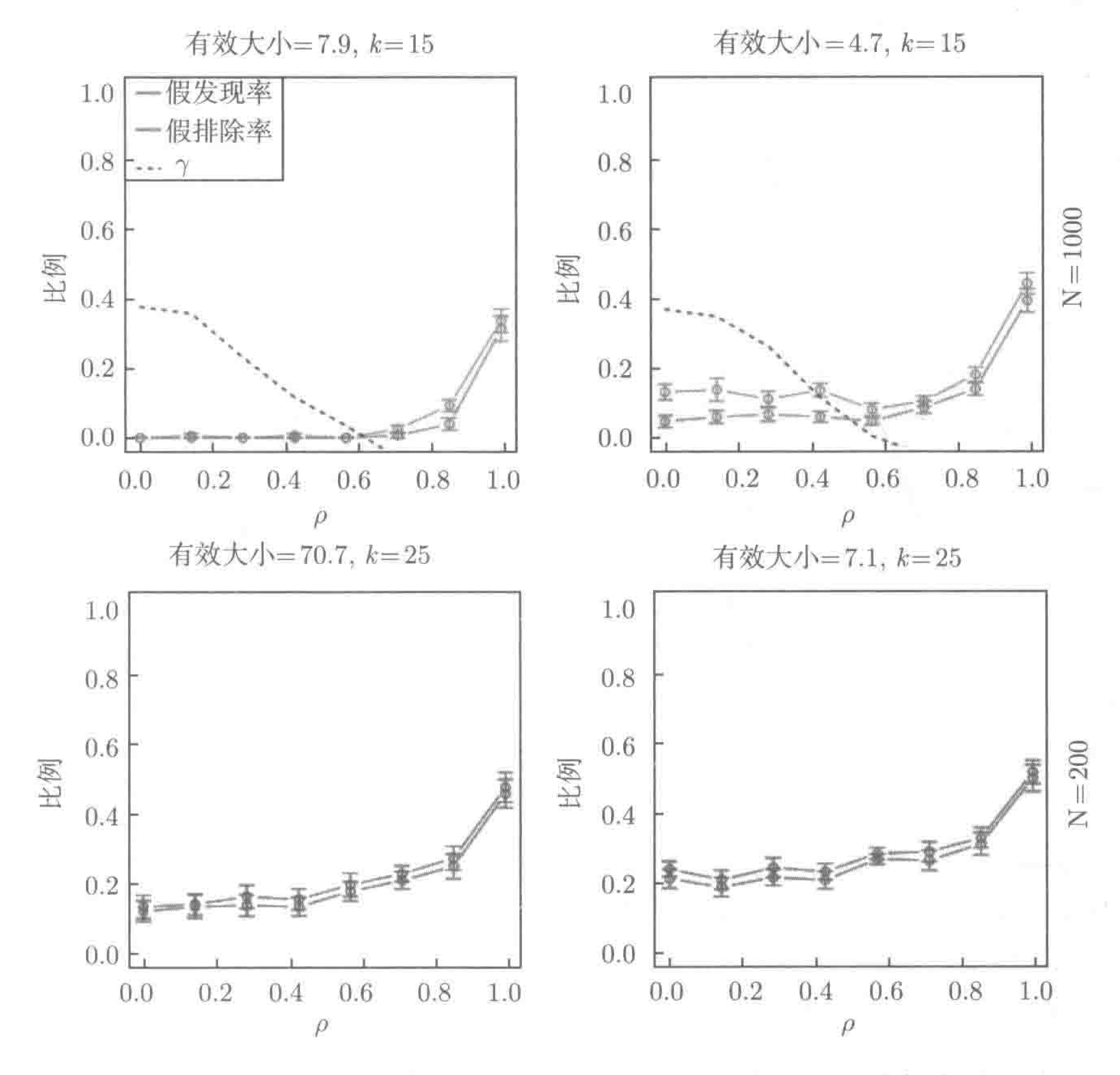


图 11-6 500 个变量下仿真实验中的平均假发现率和假排除率（±1 个标准差）。上面一行为 $N=1000$，S 的大小为 $k=15$。下面一行为 $N=200$，子集大小为 $k=25$。有效大小是真实系数的强度，用绝对 Z 统计量来衡量。总的来说，当 γ 有利时，信号强烈，可以恢复得很好（见左上图）。其他情况都有问题

图 11-6 的上面一行是结果。在左上图中（最好情况），在 ρ 大于 0.6 之前，平均假发现率和假排除率都为 0。过了那点，lasso 算法开始包含假变量，排除真实变量，因为真实变量和噪声变量之间有强相关性。图中也显示互不相关性条件的 γ 值，在 $\rho = 0.6$ 附近该值位于 0 下。（因此在 ρ 小于 0.6 时条件成立。）从右上图可以看出，即使对小的 ρ，误差率也会全面增大。这里，有效大小从 7.9 降到 4.7，这就是误差率增大的原因。

在图 11-6 的下半部分，样本数目 N 降到了 200（$p < N$），而支撑集的大小 k 增大到了 25。非零回归系数的值为 5.0 和 0.5，得到的有效大小分别为 71 和 7。互不相关性条件和理论的其他假设不成立。现在错误率为 15%，并且与 ρ 无关。在这种情况下，得到的真实支撑集看起来并不真实。

11.4.2 定理 11.3 的证明

下面来讨论 lasso 中最优化的充分必要条件。ℓ_1 范数不可导（因为零处有一个尖点），这会带来一些问题。求解需要用到 ℓ_1 范数的次微分，下面进行简单的介绍（更详细的介绍见第 5 章）。给定凸函数 $f:\mathbb{R}^p \to \mathbb{R}$，$z \in \mathbb{R}^p$ 为 β 处的次梯度，用 $\boldsymbol{z} \in \partial f(\beta)$ 表示，如果当 $f(\beta) = \|\beta\|_1$ 时有

$$f(\beta + \Delta) \geqslant f(\beta) + \langle \boldsymbol{z}, \Delta \rangle, \quad \Delta \in \mathbb{R}^p$$

则当且仅当 $\boldsymbol{z}_j = \text{sgn}(\beta_j)$ 有 $\boldsymbol{z} \in \partial \|\beta\|_1$，其中 $j = 1, 2, \ldots, p$，且允许 sgn(0) 为区间 $[-1, 1]$ 的任意一个数。在 lasso 问题中，如果 $\widehat{\beta}$ 是一个最小值，且 $\widehat{\boldsymbol{z}} \in \partial\|\widehat{\beta}\|_1$，则称 $(\widehat{\beta}, \widehat{\boldsymbol{z}}) \in \mathbb{R}^p \times \mathbb{R}^p$ 是原始–对偶（primal-dual）的最优解。这样的对必须满足次梯度为 0 的条件，即

$$\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) + \lambda_N \hat{\boldsymbol{z}} = 0 \tag{11.32}$$

在这种不可微情况下，这与梯度为 0 的条件类似。

定理 11.3 的证明基于一个构造性过程，即**原–对偶证据**（Primal-Dual Witness，PDW）方法。依照这个方法，可以得到一对 $(\widehat{\beta}, \widehat{\boldsymbol{z}}) \in \mathbb{R}^p \times \mathbb{R}^p$ 原始–对偶最优解，这样就证明了一个事实：lasso 有唯一最优解，且有正确的符号支持。这里用 $S = \text{supp}(\beta^*)$ 来表示 β^* 的支撑集，该方法包含以下步骤。

原始–对偶证据（PDW）法的构建

(1) 设 $\widehat{\beta}_{S^c} = 0$。

(2) 通过求解 k 维“最佳情况”子问题

$$\hat{\beta}_S \in \arg \min_{\beta_S \in \mathbb{R}^k} \left\{ \frac{1}{2N} \|\boldsymbol{y} - \boldsymbol{X}_S \beta_S\|_2^2 + \lambda_N \|\beta_S\|_1 \right\} \tag{11.33}$$

来确定 $(\widehat{\beta}_S, \widehat{\boldsymbol{z}}_S)$，其中 $\widehat{\boldsymbol{z}}_S$ 是次微分 $\partial\|\widehat{\beta}_S\|_1$ 的元素，满足关系 $\frac{1}{N}\boldsymbol{X}_S^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}_S\widehat{\beta}_S) + \lambda_N \widehat{\boldsymbol{z}}_S = 0$。

(3) 通过次梯度为 0 的式（11.32）求解 $\widehat{\boldsymbol{z}}_{S^c}$，检查**严格对偶可行**（strict dual feasibility）条件 $\|\widehat{\boldsymbol{z}}_{S^c}\|_\infty < 1$ 是否成立。

这个过程并不能真正用来求解 lasso 问题，因为事先假设了真实支撑集。这只能证明 lasso 的变量选择一致性。注意，子向量 $\widehat{\beta}_{S^c}$ 在步骤 (1) 中已确定，而余下的 3 个子向量在步骤 (2) 和步骤 (3) 中确定。在这样的构造下，子向量 $\widehat{\beta}_S$、$\widehat{\boldsymbol{z}}_S$ 和 $\widehat{\boldsymbol{z}}_{S^c}$ 满足次梯度为 0 的条件（11.32）。如果在步骤 (3) 中构造的向量 $\widehat{\boldsymbol{z}}_{S^c}$ 满足严格对偶可行条件，则可以认为 PDW 构建成功了。下面的结果显示了这种成功将作为 lasso 的一个见证。

引理 11.2 如果 PDW 构建成功，则在较低特征值下界条件（11.29）下，向量 $(\widehat{\boldsymbol{\beta}}_S, 0) \in \mathbb{R}^p$ 是正则化 lasso（11.3）的唯一最优解。

证明 PDW 构建成功，则 $\widehat{\beta}_S = (\widehat{\beta}_S, 0)$ 是最优解，次梯度向量 $\widehat{\boldsymbol{z}} \in \mathbb{R}^p$ 满足 $\|\widehat{\boldsymbol{z}}_{S^c}\|_\infty < 1$，且 $\langle \widehat{\boldsymbol{z}}, \widehat{\beta} \rangle = \|\widehat{\beta}\|_1$。现在设 $\widetilde{\beta} \in \mathbb{R}^p$ 是 lasso 的其他最优解。如果记 $F(\beta) = \dfrac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$，并使 $F(\widehat{\beta}) + \lambda_N \langle \widehat{\boldsymbol{z}}, \widehat{\beta} \rangle = F(\widetilde{\beta}) + \lambda_N \|\widetilde{\beta}\|_1$，则有

$$F(\hat{\beta}) - \lambda_N \langle \hat{\boldsymbol{z}}, \tilde{\beta} - \hat{\beta} \rangle = F(\tilde{\beta}) + \lambda_N (\|\tilde{\beta}\|_1 - \langle \hat{\boldsymbol{z}}, \tilde{\beta} \rangle)$$

由次梯度为 0 条件（11.32）可知：$\lambda_N \widehat{\boldsymbol{z}} = -\nabla F(\widehat{\beta})$，于是

$$F(\hat{\beta}) + \langle \nabla F(\hat{\beta}), \tilde{\beta} - \hat{\beta} \rangle - F(\tilde{\beta}) = \lambda_N (\|\tilde{\beta}\|_1 - \langle \hat{\boldsymbol{z}}, \tilde{\beta} \rangle)$$

因为 F 是凸的，所以公式左边为负，因此必须有 $\|\widetilde{\beta}\|_1 \leqslant \langle \widehat{\boldsymbol{z}}, \widetilde{\beta} \rangle$。由 ℓ_1 和 $\ell\infty$ 范数的 Hölder 不等式可以得到上界 $\langle \widehat{\boldsymbol{z}}, \widetilde{\beta} \rangle \leqslant \|\widehat{\boldsymbol{z}}\|_\infty \|\widetilde{\beta}\|_1$。这两个不等式同时成立意味着 $\|\widetilde{\beta}\|_1 = \langle \widehat{\boldsymbol{z}}, \widetilde{\beta} \rangle$。因为 $\|\widehat{\boldsymbol{z}}_{S^c}\|_\infty < 1$，所以对所有 $j \in S^c$，若 $\widetilde{\beta}_j = 0$ 则这个等式成立。

因此，任意最优解只有在 S 上才成立，并可通过求解“最佳情况”子问题（11.33）得到。给定特征值下界（11.29），这个子问题是严格凸的，因此有唯一最小值。

基于引理 11.2，要证明定理 11.3 的 (a) 和 (b)，则证明步骤 (3) 中构造的子向量 $\widehat{\boldsymbol{z}}_{S^c}$ 满足严格对偶可行条件 $\|\widehat{\boldsymbol{z}}_{S^c}\|_\infty < 1$ 即可。

确立严格对偶可行。 先深入研究步骤 (3) 中构造的子向量 $\widehat{\boldsymbol{z}}_{S^c}$。利用 $\widehat{\beta}_{S^c} = \beta^*_{S^c} = 0$，并以块矩阵的形式写出次梯度为 0 的条件（11.32），可以得到

$$\frac{1}{N}\begin{bmatrix} \boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S & \boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_{S^c} \\ \boldsymbol{X}_{S^c}^{\mathrm{T}}\boldsymbol{X}_S & \boldsymbol{X}_{S^c}^{\mathrm{T}}\boldsymbol{X}_{S^c} \end{bmatrix}\begin{bmatrix} \hat{\beta}_S - \beta^*_S \\ 0 \end{bmatrix} + \frac{1}{N}\begin{bmatrix} \boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{w} \\ \boldsymbol{X}_{S^c}^{\mathrm{T}}\boldsymbol{w} \end{bmatrix} + \lambda_N \begin{bmatrix} \hat{\boldsymbol{z}}_S \\ \hat{\boldsymbol{z}}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{11.34}$$

求解向量 $\widehat{\boldsymbol{z}}_{S^c} \in \mathbb{R}^{p-k}$ 得到

$$\hat{\boldsymbol{z}}_{S^c} = \frac{1}{\lambda_N}\left\{\frac{\boldsymbol{X}_{S^c}^{\mathrm{T}}\boldsymbol{X}_S}{N}(\hat{\beta}_S - \beta^*_S) + \frac{\boldsymbol{X}_{S^c}^{\mathrm{T}}\boldsymbol{w}}{N}\right\} \tag{11.35}$$

同样，假定 $\boldsymbol{X}_S^T\boldsymbol{X}_S$ 可逆，以便求解差 $\widehat{\beta}_S-\beta_S^*$。

$$\hat{\beta}_S-\beta_S^*=\underbrace{-\left(\frac{\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S}{N}\right)^{-1}\frac{\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{w}}{N}-\lambda_N\left(\frac{\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S}{N}\right)^{-1}\mathrm{sgn}(\beta_S^*)}_{U_S} \tag{11.36}$$

将这一项代回式（11.35），然后化简得到

$$\hat{\boldsymbol{z}}_{S^c}=\underbrace{\boldsymbol{X}_{S^c}^{\mathrm{T}}\boldsymbol{X}_S(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S)^{-1}\mathrm{sgn}(\beta_S^*)}_{\mu}+\underbrace{\boldsymbol{X}_{S^c}^{\mathrm{T}}\left[\boldsymbol{I}-\boldsymbol{X}_S(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^{\mathrm{T}}\right]\left(\frac{\boldsymbol{w}}{\lambda_N N}\right)}_{V_{S^c}}$$

再运用三角不等式，则有

$$\|\hat{\boldsymbol{z}}_{S^c}\|_\infty\leqslant\|\mu\|_\infty+\|V_{S^c}\|_\infty$$

注意，向量 $\mu\in\mathbb{R}^{p-k}$ 是一个定量，而且由互不相关性条件（11.27）可以得到 $\|\mu\|_\infty\leqslant 1-\gamma$。剩余量 $V_{S^c}\in\mathbb{R}^{p-k}$ 是一个零均值高斯随机向量，需要证明 $\|V_{S^c}\|_\infty<\gamma$ 有很高的概率。

对任意 $j\in S^c$，考虑随机变量

$$V_j:=\boldsymbol{X}_j^{\mathrm{T}}\underbrace{\left[\boldsymbol{I}-\boldsymbol{X}_S(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^{\mathrm{T}}\right]}_{\Pi_{S^\perp}(\boldsymbol{X})}\left(\frac{\boldsymbol{w}}{\lambda_N N}\right)$$

注意，矩阵 $\prod_{S^\perp}(\boldsymbol{X})$ 是一个正交投影矩阵，采用列归一化条件（11.28）可以得出结论：任意 V_j 为零均值，方差最大为 $\sigma^2K_{\mathrm{clm}}^2/(\lambda_N^2N)$。因此，结合高斯截尾界及一致界，就有

$$\mathbb{P}\left[\|V_{S^c}\|_\infty\geqslant\gamma/2\right]\leqslant 2(p-k)e^{-\frac{\lambda_N^2N(\gamma/2)^2}{2\sigma^2K_{\mathrm{clm}}^2}}$$

对于定理中给定的 λ_N，概率以速率 $2e^{-2\lambda_N^2N}$ 减小。

确立 ℓ_∞ 界。 下面确立式（11.36）中差向量 $U_S=\widehat{\beta}_S-\beta_S^*$ 的 ℓ_∞ 范数的界。由三角不等式可得到

$$\|U_S\|_\infty\leqslant\left\|\left(\frac{\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S}{N}\right)^{-1}\frac{\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{w}}{N}\right\|_\infty+\left\|\left(\frac{\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S}{N}\right)^{-1}\right\|_\infty\lambda_N \tag{11.37}$$

为了方便，这里为每项乘以 $\dfrac{1}{N}$。第二项是一个确定量，因此保留它以便限制第一项。对于任意 $i=1,\ldots,k$，考虑随机变量

$$\boldsymbol{Z}_i:=\mathrm{e}_i^{\mathrm{T}}\left(\frac{1}{N}\boldsymbol{X}_S^T\boldsymbol{X}_S\right)^{-1}\frac{1}{N}\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{w}$$

因为 $\boldsymbol{w}$ 的元素服从独立同分布 $N\left(0,\sigma^2\right)$，变量 $\boldsymbol{Z}_i$ 是零均值高斯变量，所以方差最大为

$$\frac{\sigma^2}{N}\left\|\left(\frac{1}{N}\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S\right)^{-1}\right\|_2 \leqslant \frac{\sigma^2}{C_{\min}N}$$

这里使用了特征值条件（11.29）。因此，再一次结合高斯截尾边界和联合边界，就得到

$$\mathbb{P}\left[\|U_S\|_\infty > t\right] \leqslant 2\mathrm{e}^{-\frac{t^2 C_{\min} N}{2\sigma^2}+\log k}$$

设 $t=4\sigma\lambda_N/\sqrt{C_{\min}}$，并选择使 $8N\lambda_N^2 \geqslant \log p \geqslant \log k$ 成立的 λ_N。将这些结合在一起，可得出至少有 $1-2e^{-c_2\lambda_N^2 N}$ 的概率使 $\|U_S\|_\infty = 4\sigma\lambda_N/\sqrt{C_{\min}}$。最后结论为

$$\left\|\hat{\beta}_S-\beta_S^*\right\|_\infty \leqslant \lambda_N\left[\frac{4\sigma}{\sqrt{C_{\min}}}+\left\|(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S/N)^{-1}\right\|_\infty\right]$$

如同之前声明的那样，这个不等式成立的概率大于 $1-2e^{-c_2\lambda_N^2 N}$。

11.5 超越基础 lasso

本章仅讨论基本的 lasso，即将最小二乘损失函数与 ℓ_1 范数正则相结合。但是，该损失函数能直接扩展到更一般的损失函数，包括逻辑斯蒂回归和其他类型的广义线性函数，也可扩展成不同形式正则化，包括组 lasso、核范数及其他形式的结构化正则子。这里只介绍基本的, 文献注释会更加详细地进行介绍。

考虑目标函数

$$F(\beta)=\frac{1}{N}\sum_{i=1}^{N} f(\beta;\boldsymbol{z}_i) \tag{11.38}$$

其中函数 $\beta\to g\left(\beta;\boldsymbol{z}_i\right)$ 度量参数向量 $\beta\in\mathbb{R}^p$ 对样本 $\boldsymbol{z}_i$ 的拟合情况。在回归问题中，样本的形式为 $\boldsymbol{z}_i=(x_i,y_i)\in\mathbb{R}^p\times\mathbb{R}$，而在图 lasso 问题中，样本对应向量 $z_i=x_i\in\mathbb{R}^p$。设 $\varPhi:\mathbb{R}^p\to\mathbb{R}$ 表示一个正则子，则有

$$\hat{\beta}\in\arg\min_{\beta\in\Omega}\{F(\beta)+\lambda_N\Phi(\beta)\} \tag{11.39}$$

这里可将 $\widehat{\beta}$ 当成是确定向量 β^* 的估计，这个估计是通过最小化总体目标函数 $\overline{F}\left(\beta\right):=\mathbb{E}\left[f\left(\beta;\boldsymbol{Z}\right)\right]$ 得到的。

将之前的讨论放入当前情形中，则 lasso 是这个广义 M 估计的一个特例，取

$$f(\beta;x_i,y_i)=\frac{1}{2}\left(y_i-\langle x_i,\beta\rangle\right)^2,\quad \varPhi(\beta)=\|\beta\|_1$$

并在 $\boldsymbol{\Omega}=\mathbb{R}^p$ 上进行最优化，就可得到 lasso。在随机设计的情况下，协变量 $x_i\sim N\left(0,\boldsymbol{\Sigma}\right)$，线性回归的总体目标函数可写为 $\overline{F}(\beta)=\frac{1}{2}(\beta-\beta^*)^T\boldsymbol{\Sigma}(\beta-\beta^*)+\frac{1}{2}\sigma^2$。

对于广义 M 估计 (11.39)，这里的目标为分析误差 $\|\widehat{\beta}-\beta^*\|_2$ 提供了一些思路。当 $N<p$ 时，目标函数 (11.38) 不可能是强凸的。事实上，假设其二阶可微，则海森矩阵是 p 维中 N 个矩阵之和，所以必定是秩退化的。如前面所介绍的，受限特征值条件是损失函数和正则子的广义条件的特例，即限制强凸性。具体而言，给定集合 $\mathcal{C}\in\mathbb{R}^p$，如果存在参数 $\gamma>0$，可微函数 F 在 $\mathcal{C}$ 上满足限制强凸性，即

$$F(\beta^*+\nu)-F(\beta^*)-\langle\nabla F(\beta^*),\nu\rangle\geqslant\gamma\|\nu\|_2^2,\nu\in\mathcal{C} \tag{11.40}$$

那么当 F 二阶可微，下界等价于在 β^* 的邻域上限制海森矩阵，如式 (11.9) 中的定义 (详见习题 11.6)。因此，在最小二乘问题的特例上，限制强凸性等价于受限特征值条件。

任意类型的集合 $\mathcal{C}$ 都能保证这种形式的条件成立吗？因为这里的最终目标是控制误差向量 $\widehat{\nu}=\widehat{\beta}-\beta^*$，所以只需要确保子集 $\mathcal{C}$ 上强凸性成立，即保证 (在数据集上以较高的概率) 含有误差向量。对于满足**分解性** (decomposability) 的正则子，存在这样子集，将 ℓ_1 范数的基本特性推广到更广泛的正则子族上。分解性定义在参数集 Ω 的子空间 $\mathcal{M}$ 上，这意味着在最优 β^* 上描述结构，其正交补集 $\mathcal{M}^\perp$ 对应于与模型结构不相关的不利扰动。基于这样的观点，如果

$$\Phi(\beta+\theta)=\Phi(\beta)+\Phi(\theta),\ \text{对所有}(\beta,\theta)\in\mathcal{M}\times\mathcal{M}^\perp \tag{11.41}$$

则正则子 Φ 是关于 $\mathcal{M}$ 可分解的。对于 ℓ_1 范数，模型子空间可简化成在固定集 S 上有支撑的所有向量的集合的，而正交补集 $\mathcal{M}^\perp$ 包含由补集 S^c 支撑的向量。分解关系式 (11.41) 沿用了 ℓ_1 范数的坐标系特性。选择合适的子空间，很多其他正则子也可分解，包括加权 lasso、组 lasso、叠加组 lasso 惩罚，以及低秩矩阵的核范数，详见参考文献注释。

参考文献注释

Knight and Fu (2000) 推导了维数 p 固定时 lasso 的渐近理论和相关估计，并在分析中提出不相关性条件 (11.27)。Greenshtein and Ritov (2004) 首次提供了 lasso 的高维分析，特别是在 $p\gg N$ 情况下提供了预测误差界。无代表性或互不相关性条件 (11.27) 由 Fuchs (2004) 和 Tropp (2006) 在信号处理领域提出，而 Meinshausen and Bühlmann (2006) 和 Zhao and Yu (2006) 在统计学中也曾独立提出。受限特征值的概念由 Bickel, Ritov and Tsybakov (2009) 提出。这是一个比第 10 章中受限等距性弱的限制条件。van de Geer and Bühlmann (2009) 为了证明 lasso 上的估计误差界，将它与其他相关条件进行比较。Candès and Tao (2007) 定义并给出了 “Dantzig 选择器” 理论，这是一个类似于 lasso 的问题。Raskutti,

Wainwright and Yu（2010）证明了对于多种随机高斯设计矩阵，RE 条件以很高的概率成立；Rudelson and Zhou（2013）对子高斯设计矩阵的扩展进行了讨论。

定理 11.1 的证明基于 Bickel et al.（2009），而 Negahban et al.（2012）推导了 ℓ_q 稀疏向量上的 lasso 误差界（11.24）。这些证明中用到的基本不等式在 M 估计（van de Geer 2000）的分析中较为常见。Raskutti, Wainwright and Yu（2011）分析了 ℓ_q 球上回归的极大极小率，得到 ℓ_2 误差和预测误差的比率。定理 11.2(a) 由 Bunea, Tsybakov and Wegkamp（2007）证明，而定理 11.2(b) 由 Bickel et al.（2009）证明。为达到定理 11.2(b) 中给定的"快速率"，任意多项式时间的方法都需要受限特征值条件，正如 Zhang, Wainwright and Jordan（2014）的结论。在复杂理论的标准假设下，这篇文献证明了没有多项式时间算法在不加 RE 条件的情况下能够达到快速率。

定理 11.3 和原始–对偶证据由 Wainwright（2009）证明。同一文献还对设计矩阵的高斯组合确立了尖（sharp）阈值结果，特别是在样本大小上给出了上下界，这影响了在计算支撑集上的成功或失败。引理 11.2 的证明由 Caramanis（2010）提供。PDW 方法已经运用在了一系列其他问题中，包括组 lasso 问题分析（Obozinski et al. 2011 和 Wang, Liang and Xing 2013）和相关松弛（Jalali, Ravikumar, Sanghavi and Ruan 2010 和 Negahban and Wainwright 2011b），图 lasso（Ravikumar et al. 2011），以及带隐变量的高斯图选择方法（Chandrasekaran et al. 2012）。Lee, Sun and Taylor（2013）为 M 估计的更广泛类提供了 PDW 方法的广义公式。

正如 11.5 节谈到的，本章的分析可以拓展到 M 估计的更广泛的类，即基于可分解正则子的那一类。Negahban et al.（2012）提供了一个分析这类 M 估计的估计误差 $\|\widehat{\beta}-\beta^*\|_2$ 的一般框架，即两个主要部分为损失函数的限制强凸性和正则子的分解性。

习 题

习题 11.1 对于给定的 $q \in (0,1]$，将式（11.7）中定义的集合 $\mathbb{B}_q(R_q)$ 作为软稀疏性模型。

(a) 一个相关目标是弱 ℓ_q 球，参数为 (C,α)，给定

$$\mathbb{B}_{w(\alpha)}(C) := \left\{\theta \in \mathbb{R}^p \,|\, |\theta|_{(j)} \leqslant Cj^{-\alpha},\ j=1,\dots,p\right\} \tag{11.42a}$$

这里 $|\theta|_{(j)}$ 表示 θ 绝对值的有序统计量，按最大到最小进行排序（所以 $|\theta|_{(1)} = \max\limits_{j=1,2,\ldots,p} |\theta_j|$，$|\theta|_{(p)} = \min\limits_{j=1,2,\ldots,p} |\theta_j|$。）对于任意的 $\alpha > 1/q$，求证：存在一个依赖于 (C,α) 的半径 R_q，使 $\mathbb{B}_{w(\alpha)}(C) \subseteq \mathbb{B}_q(R_q)$。

(b) 对给定的整数 $k \in \{1, 2, \ldots, p\}$，向量 $\theta^* \in \mathbb{R}^p$ 的最佳 k 项近似为

$$\prod_k (\theta^*) := \arg \min_{\|\theta\|_0 \leqslant k} \|\theta - \theta^*\|_2^2 \tag{11.42b}$$

给出 $\Pi_k (\theta^*)$ 的闭合解的表达式。

(c) 对于 $q \in (0, 1]$，$\theta^* \subseteq \mathbb{B}_q (R_q)$，求证：最佳 k 项近似满足

$$\|\prod_k(\theta^*) - \theta^*\|_2^2 \leqslant (R_q)^{2/q} \left(\frac{1}{k}\right)^{\frac{2}{q}-1} \tag{11.42c}$$

习题 11.2 本习题分析 lasso 的一个替代版本，即估计子

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{，使得} \quad \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 \leqslant C \tag{11.43}$$

其中常数 $C > 0$ 是一个用户可选参数。这种形式的 lasso 常称为**松弛基追踪**。

(a) 假设选定 C，使 β^* 对凸问题可行。求证：误差向量 $\widehat{\nu} = \widehat{\beta} - \beta^*$ 必定满足锥约束 $\|\widehat{\nu}_{S^c}\|_1 \leqslant \|\widehat{\nu}_S\|_1$。

(b) 假设线性观测模型 $\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{w}$，求证：$\widehat{\nu}$ 满足基本不等式

$$\frac{\|\boldsymbol{X}\hat{\nu}\|_2^2}{N} \leqslant 2\frac{\left\|\boldsymbol{X}^T\boldsymbol{w}\right\|_\infty}{N} \|\hat{\nu}\|_1 + \left\{C - \frac{\|\boldsymbol{w}\|_2^2}{N}\right\}$$

(c) 假设存在 $\boldsymbol{X}$ 上的一个 γ-RE 条件，用 (b) 确立 ℓ_2 误差 $\|\widehat{\beta} - \beta^*\|_2$ 的界。

习题 11.3 本习题介绍界（11.24）的证明。具体而言，这里要证明，如果 $\lambda_N \geqslant \frac{\|\boldsymbol{X}^T\boldsymbol{w}\|_\infty}{N}$，且 $\beta^* \in \mathbb{B}_q (R_q)$，则拉格朗日 lasso 误差满足界

$$\|\hat{\beta} - \beta^*\|_2^2 \leqslant cR_q\lambda_N^{1-q/2} \tag{11.44a}$$

(a) 求证：误差向量 $\widehat{\nu}$ 对任意子集 $S \subseteq \{1, 2, ..., p\}$ 及其补集满足“类锥”约束

$$\|\hat{\nu}_{S^c}\|_1 \leqslant 3\|\hat{\nu}_S\|_1 + \|\beta_{S^c}^*\|_1 \tag{11.44b}$$

由此推广引理 11.1。

(b) 假设 $\boldsymbol{X}$ 在所有向量上满足 γ-RE 条件，而这些向量满足类锥条件（11.44b）。求证：

$$\|\hat{\nu}\|_2^2 \leqslant \lambda_N \{4\|\hat{\nu}_S\|_1 + \|\beta_{S^c}^*\|_1\}$$

对任意有序子集 S 有效。

(c) 优化 S 的选择，得到界（11.44a）。

习题 11.4　设有随机设计矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，其每一行 $x_i \in \mathbb{R}^p$ 独立同分布地从 $\mathcal{N}(0, \boldsymbol{\Sigma})$ 中得到，其中协方差矩阵 $\boldsymbol{\Sigma}$ 严格正定。求证：对一个足够大的常数 c，样本数的下界为 $N > c|S|^2 \log p$ 时，γ-RE 条件在集合 $\mathcal{C}(S;\alpha)$ 上以高概率成立。（注意：样本数的规模不是最优的，可以用一个更好的参数将 $|S|^2$ 降到 $|S|$。）

习题 11.5　考虑随机设计矩阵 $\boldsymbol{X} \in \mathbb{R}^{N\times p}$，它的各项服从独立同分布 $N(0,1)$。本习题要证明：只要有一个足够大的常数 c 使得 $N > ck\log p$，互不相关性条件（11.27）就会以高概率成立。（**提示**：对 $N > 4k$，$\varepsilon = \{\lambda_{\min}(\frac{\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S}{N}) \geqslant \frac{1}{4}\}$ 以高概率成立。）

(a) 求证：

$$\gamma = 1 - \max_{j\in S^c} \max_{z\in\{-1,+1\}^k} \underbrace{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{X}_S(\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{X}_S)^{-1}z}_{V_{j,z}}$$

(b) 回忆事件 ε，求证：存在数值常数 $c_0 > 0$，使得对任意固定索引 $j \in S^c$ 和向量 $\boldsymbol{z} \in \{-1,+1\}^k$，

$$\mathbb{P}[V_{j,z} \geqslant t] \leqslant \mathrm{e}^{-c_0\frac{Nt^2}{k}} + \mathbb{P}[\varepsilon^c]，\quad t > 0$$

(c) 运用 (b) 完成证明。

习题 11.6　对于二阶可微函数 $F:\mathbb{R}^p \to \mathbb{R}$ 和集合 $\mathcal{C} \subset \mathbb{R}^p$，对固定参数 β^* 的邻域内的所有 β，

$$\frac{\nabla^2 F(\beta)}{\|\nu\|_2^2} \geqslant \gamma\|\nu\|_2^2，\quad \nu \in \mathcal{C}$$

一致。求证：RSC 条件（11.40）成立。

参考文献

Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P. and Tandon, R. (2014), Learning sparsely used overcomplete dictionaries via alternating minimization, *Journal of Machine Learning Research Workshop* **35**, 123–137.

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012*a*), Fast global convergence of gradient methods for high-dimensional statistical recovery, *Annals of Statistics* **40**(5), 2452–2482.

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012*b*), Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions, *Annals of Statistics* **40**(2), 1171–1197.

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Pwellm, J., Marti, G., Moore, T., Hudsom, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R., Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. and Staudt, L. (2000), Identification of molecularly and clinically distinct subtypes of diffuse large b cell lymphoma by gene expression profiling, *Nature* **403**, 503–511.

Alliney, S. and Ruzinsky, S. (1994), An algorithm for the minimization of mixed L1 and L2 norms with application to Bayesian estimation, *Transactions on Signal Processing* **42**(3), 618–627.

Amini, A. A. and Wainwright, M. J. (2009), High-dimensional analysis of semdefinite relaxations for sparse principal component analysis, *Annals of Statistics* **5B**, 2877–2921.

Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis, 3rd ed.*, Wiley, New York.

Antoniadis, A. (2007), Wavelet methods in statistics: Some recent developments and their applications, *Statistics Surveys* **1**, 16–55.

Bach, F. (2008), Consistency of trace norm minimization, *Journal of Machine Learning Research* **9**, 1019–1048.

Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012), Optimization with sparsity-inducing penalties, *Foundations and Trends in Machine Learning* **4**(1), 1–106.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008), Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, *Journal of Machine Learning Research* **9**, 485–516.

Baraniuk, R. G., Davenport, M. A., DeVore, R. A. and Wakin, M. B. (2008), A simple proof of the restricted isometry property for random matrices, *Constructive Approximation* **28**(3), 253–263.

Barlow, R. E., Bartholomew, D., Bremner, J. M. and Brunk, H. D. (1972), *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*, Wiley, New York.

Beck, A. and Teboulle, M. (2009), A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* **2**, 183–202.

Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B*. **85**, 289–300.

Bennett, J. and Lanning, S. (2007), The netflix prize, in *Proceedings of KDD Cup and Workshop in conjunction with KDD*.

Bento, J. and Montanari, A. (2009), Which graphical models are difficulty to learn?, in *Advances in Neural Information Processing Systems (NIPS Conference Proceedings)*.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013), Valid post-selection inference, *Annals of Statistics* **41**(2), 802–837.

Berthet, Q. and Rigollet, P. (2013), Computational lower bounds for sparse PCA, Technical report, Princeton University. arxiv1304.0828.

Bertsekas, D. (1999), *Nonlinear Programming*, Athena Scientific, Belmont MA.

Bertsekas, D. (2003), *Convex Analysis and Optimization*, Athena Scientific, Belmont MA.

Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society Series B* **36**, 192–236.

Besag, J. (1975), Statistical analysis of non-lattice data, *The Statistician* **24**(3), 179–195.

Bickel, P. J. and Levina, E. (2008), Covariance regularization by thresholding, *Annals of Statistics* **36**(6), 2577–2604.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009), Simultaneous analysis of Lasso and Dantzig selector, *Annals of Statistics* **37**(4), 1705–1732.

Bien, J., Taylor, J. and Tibshirani, R. (2013), A Lasso for hierarchical interactions, *Annals of Statistics* **42**(3), 1111–1141.

Bien, J. and Tibshirani, R. (2011), Sparse estimation of a covariance matrix, *Biometrika* **98**(4), 807–820.

Birnbaum, A., Johnstone, I., Nadler, B. and Paul, D. (2013), Minimax bounds for sparse PCA with noisy high-dimensional data, *Annals of Statistics* **41**(3), 1055–1084.

Boser, B., Guyon, I. and Vapnik, V. (1992), A training algorithm for optimal

margin classifiers, in *Proceedings of the Annual Conference on Learning Theory (COLT)*, Philadelphia, Pa.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011), Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends in Machine Learning* **3**(1), 1–124.

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press, Cambridge, UK.

Breiman, L. (1995), Better subset selection using the nonnegative garrote, *Technometrics* **37**, 738–754.

Breiman, L. and Ihaka, R. (1984), Nonlinear discriminant analysis via scaling and ACE, Technical report, University of California, Berkeley.

Bühlmann, P. (2013), Statistical significance in high-dimensional linear models, *Bernoulli* **19**(4), 1212–1242.

Bühlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, New York.

Bunea, F., She, Y. and Wegkamp, M. (2011), Optimal selection of reduced rank estimators of high-dimensional matrices, **39**(2), 1282–1309.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007), Sparsity oracle inequalities for the Lasso, *Electronic Journal of Statistics* pp. 169–194.

Burge, C. and Karlin, S. (1977), Prediction of complete gene structures in human genomic DNA, *Journal of Molecular Biology* **268**, 78–94.

Butte, A., Tamayo, P., Slonim, D., Golub, T. and Kohane, I. (2000), Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, *Proceedings of the National Academy of Sciences* pp. 12182–12186.

Candès, E., Li, X., Ma, Y. and Wright, J. (2011), Robust Principal Component Analysis?, *Journal of the Association for Computing Machinery* **58**, 11:1–11:37.

Candès, E. and Plan, Y. (2010), Matrix completion with noise, *Proceedings of the IEEE* **98**(6), 925–936.

Candès, E. and Recht, B. (2009), Exact matrix completion via convex optimization, *Foundation for Computational Mathematics* **9**(6), 717–772.

Candès, E., Romberg, J. K. and Tao, T. (2006), Stable signal recovery from incomplete and inaccurate measurements, *Communications on Pure and Applied Mathematics* **59**(8), 1207–1223.

Candès, E. and Tao, T. (2005), Decoding by linear programming, *IEEE Transactions on Information Theory* **51**(12), 4203–4215.

Candès, E. and Tao, T. (2007), The Dantzig selector: Statistical estimation when p is much larger than n, *Annals of Statistics* **35**(6), 2313–2351.

Candès, E. and Wakin, M. (2008), An introduction to compressive sampling, *Signal Processing Magazine, IEEE* **25**(2), 21–30.

Caramanis, C. (2010), 'Personal communication'.

Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012), Latent variable graphical model selection via convex optimization, *Annals of Statistics* **40**(4), 1935–1967.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. and Willsky, A. S. (2011), Rank-sparsity incoherence for matrix decomposition, *SIAM Journal on Optimization* **21**, 572–596.

Chaudhuri, S., Drton, M. and Richardson, T. S. (2007), Estimation of a covariance matrix with zeros, *Biometrika* pp. 1–18.

Chen, S., Donoho, D. and Saunders, M. (1998), Atomic decomposition by basis pursuit, *SIAM Journal of Scientific Computing* **20**(1), 33–61.

Cheng, J., Levina, E. and Zhu, J. (2013), High-dimensional Mixed Graphical Models, *arXiv:1304.2810* .

Chi, E. C. and Lange, K. (2014), Splitting methods for convex clustering, *Journal of Computational and Graphical Statistics (online access)* .

Choi, Y., Taylor, J. and Tibshirani, R. (2014), Selecting the number of principal components: estimation of the true rank of a noisy matrix. arXiv:1410.8260.

Clemmensen, L. (2012), *sparseLDA: Sparse Discriminant Analysis*. R package version 0.1-6.
URL: *http://CRAN.R-project.org/package=sparseLDA*

Clemmensen, L., Hastie, T., Witten, D. and Ersboll, B. (2011), Sparse discriminant analysis, *Technometrics* **53**, 406–413.

Clifford, P. (1990), Markov random fields in statistics, in G. Grimmett and D. J. A. Welsh, eds, *Disorder in physical systems*, Oxford Science Publications.

Cohen, A., Dahmen, W. and DeVore, R. A. (2009), Compressed sensing and best k-term approximation, *Journal of the American Mathematical Society* **22**(1), 211–231.

Cox, D. and Wermuth, N. (1996), *Multivariate Dependencies: Models, Analysis and Interpretation*, Chapman & Hall, London.

d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008), First order methods for sparse covariance selection, *SIAM Journal on Matrix Analysis and its Applications* **30**(1), 55–66.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007), A direct formulation for sparse PCA using semidefinite programming, *SIAM Review* **49**(3), 434–448.

Davidson, K. R. and Szarek, S. J. (2001), Local operator theory, random matrices, and Banach spaces, in *Handbook of Banach Spaces*, Vol. 1, Elsevier, Amsterdam, NL, pp. 317–336.

De Leeuw, J. (1994), Block-relaxation algorithms in statistics, in H. Bock,

W. Lenski and M. M. Richter, eds, *Information Systems and Data Analysis*, Springer-Verlag, Berlin.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. and West, M. (2004), Sparse graphical models for exploring gene expression data, *Journal of Multivariate Analysis* **90**(1), 196 – 212.

Donoho, D. (2006), Compressed sensing, *IEEE Transactions on Information Theory* **52**(4), 1289–1306.

Donoho, D. and Huo, X. (2001), Uncertainty principles and ideal atomic decomposition, *IEEE Trans. Info Theory* **47**(7), 2845–2862.

Donoho, D. and Johnstone, I. (1994), Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**, 425–455.

Donoho, D. and Stark, P. (1989), Uncertainty principles and signal recovery, *SIAM Journal of Applied Mathematics* **49**, 906–931.

Donoho, D. and Tanner, J. (2009), Counting faces of randomly-projected polytopes when the projection radically lowers dimension, *Journal of the American Mathematical Society* **22**(1), 1–53.

Dudoit, S., Fridlyand, J. and Speed, T. (2002), Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association* **97**(457), 77–87.

Edwards, D. (2000), *Introduction to Graphical Modelling, 2nd Edition*, Springer, New York.

Efron, B. (1979), Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**, 1–26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling plans*, Vol. 38, SIAM- CBMS-NSF Regional Conference Series in Applied Mathematics.

Efron, B. (2011), The bootstrap and Markov Chain Monte Carlo, *Journal of Biopharmaceutical Statistics* **21**(6), 1052–1062.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, London.

El Ghaoui, L., Viallon, V. and Rabbani, T. (2010), Safe feature elimination in sparse supervised learning, *Pacific journal of optimization* **6**(4), 667–698.

El Karoui, N. (2008), Operator norm consistent estimation of large-dimensional sparse covariance matrices, *Annals of Statistics* **36**(6), 2717–2756.

Elad, M. and Bruckstein, A. M. (2002), A generalized uncertainty principle and sparse representation in pairs of bases, *IEEE Transactions on Information Theory* **48**(9), 2558–2567.

Erdos, P. and Renyi, A. (1961), On a classical problem of probability theory, *Magyar Tud. Akad. Mat. Kutat Int. Kzl.* **6**, 215–220. (English and Russian summary).

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P. and Bengio, S. (2010), Why does unsupervised pre-training help deep learning?, *Journal of Machine Learning Research* **11**, 625–660.

Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fazel, M. (2002), Matrix Rank Minimization with Applications, PhD thesis, Stanford. Available online: http://faculty.washington.edu/mfazel/thesis-final.pdf.

Feuer, A. and Nemirovski, A. (2003), On sparse representation in pairs of bases, *IEEE Transactions on Information Theory* **49**(6), 1579–1581.

Field, D. (1987), Relations between the statistics of natural images and the response properties of cortical cells, *Journal of the Optical Society of America* **A4**, 2379–2394.

Fisher, M. E. (1966), On the Dimer solution of planar Ising models, *Journal of Mathematical Physics* **7**, 1776–1781.

Fithian, W., Sun, D. and Taylor, J. (2014), Optimal inference after model selection, *ArXiv e-prints* .

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007), Pathwise coordinate optimization, *Annals of Applied Statistics* **1**(2), 302–332.

Friedman, J., Hastie, T., Simon, N. and Tibshirani, R. (2015), *glmnet: Lasso and elastic-net regularized generalized linear models.* R package version 2.0.

Friedman, J., Hastie, T. and Tibshirani, R. (2008), Sparse inverse covariance estimation with the graphical Lasso, *Biostatistics* **9**, 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2010*a*), Applications of the Lasso and grouped Lasso to the estimation of sparse graphical models, Technical report, Stanford University, Statistics Department.

Friedman, J., Hastie, T. and Tibshirani, R. (2010*b*), Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**(1), 1–22.

Fuchs, J. (2000), On the application of the global matched filter to doa estimation with uniform circular arrays, in *Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference - Volume 05*, ICASSP '00, IEEE Computer Society, Washington, DC, USA, pp. 3089–3092.

Fuchs, J. (2004), Recovery of exact sparse representations in the presence of noise, in *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 533–536.

Gannaz, I. (2007), Robust estimation and wavelet thresholding in partially linear models, *Statistics and Computing* **17**(4), 293–310.

Gao, H. and Bruce, A. (1997), Waveshrink with firm shrinkage, *Statistica Sinica* **7**, 855–874.

Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Golub, G. and Loan, C. V. (1996), *Matrix Computations*, Johns Hopkins University Press, Baltimore.

Gorski, J., Pfeuffer, F. and Klamroth, K. (2007), Biconvex sets and optimization with biconvex functions: a survey and extensions, *Mathematical Methods of Operations Research* **66**(3), 373–407.

Gramacy, R. (2011), 'The monomvn package: Estimation for multivariate normal and student-t data with monotone missingness', CRAN. R package version 1.8.

Grazier G'Sell, M., Taylor, J. and Tibshirani, R. (2013), Adaptive testing for the graphical Lasso. arXiv: 1307.4765.

Grazier G'Sell, M., Wager, S., Chouldechova, A. and Tibshirani, R. (2015), Sequential selection procedures and false discovery rate control. arXiv: 1309.5352: To appear, Journal of the Royal Statistical Society Series B.

Greenshtein, E. and Ritov, Y. (2004), Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization, *Bernoulli* **10**, 971–988.

Greig, D. M., Porteous, B. T. and Seheuly, A. H. (1989), Exact maximum a posteriori estimation for binary images, *Journal of the Royal Statistical Society Series B* **51**, 271–279.

Grimmett, G. R. (1973), A theorem about random fields, *Bulletin of the London Mathematical Society* **5**, 81–84.

Gross, D. (2011), Recovering low-rank matrices from few coefficients in any basis, *IEEE Transactions on Information Theory* **57**(3), 1548–1566.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer Series in Statistics, Springer, New York, NY.

Hammersley, J. M. and Clifford, P. (1971), Markov fields on finite graphs and lattices. Unpublished.

Hastie, T., Buja, A. and Tibshirani, R. (1995), Penalized discriminant analysis, *Annals of Statistics* **23**, 73–102.

Hastie, T. and Mazumder, R. (2013), *softImpute: matrix completion via iterative soft-thresholded SVD*. R package version 1.0.
URL: *http://CRAN.R-project.org/package=softImpute*

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall, London.

Hastie, T. and Tibshirani, R. (2004), Efficient quadratic regularization for expression arrays, *Biostatistics,* **5**, 329–340.

Hastie, T., Tibshirani, R. and Buja, A. (1994), Flexible discriminant analysis by optimal scoring, *Journal of the American Statistical Association* **89**, 1255–1270.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, second edn, Springer Verlag, New York.

Hastie, T., Tibshirani, R., Narasimhan, B. and Chu, G. (2003), *pamr: Prediction Analysis for Microarrays in R.* R package version 1.54.1.
URL: *http://CRAN.R-project.org/package=pamr*

Hocking, T., Vert, J.-P., Bach, F. and Joulin, A. (2011), Clusterpath: an algorithm for clustering using convex fusion penalties., in L. Getoor and T. Scheffer, eds, *Proceedings of the Twenty-Eighth International Conference on Machine Learning (ICML)*, Omnipress, pp. 745–752.

Hoefling, H. (2010), A path algorithm for the fused Lasso signal approximator, *Journal of Computational and Graphical Statistics* **19**(4), 984–1006.

Hoefling, H. and Tibshirani, R. (2009), Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods, *Journal of Machine Learning Research* **19**, 883–906.

Horn, R. A. and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge University Press, Cambridge.

Hsu, D., Kakade, S. M. and Zhang, T. (2011), Robust matrix decomposition with sparse corruptions, *IEEE Transactions on Information Theory* **57**(11), 7221–7234.

Huang, J., Ma, S. and Zhang, C.-H. (2008), Adaptive Lasso for sparse high-dimensional regression models, *Statistica Sinica* **18**, 1603–1618.

Huang, J. and Zhang, T. (2010), The benefit of group sparsity, *The Annals of Statistics* **38**(4), 1978–2004.

Hunter, D. R. and Lange, K. (2004), A tutorial on MM algorithms, *The American Statistician* **58**(1), 30–37.

Ising, E. (1925), Beitrag zur theorie der ferromagnetismus, *Zeitschrift für Physik* **31**(1), 253–258.

Jacob, L., Obozinski, G. and Vert, J.-P. (2009), Group Lasso with overlap and graph Lasso, in *Proceeding of the 26th International Conference on Machine Learning, Montreal, Canada.*

Jalali, A., Ravikumar, P., Sanghavi, S. and Ruan, C. (2010), A dirty model for multi-task learning, in *Advances in Neural Information Processing Systems 23*, pp. 964–972.

Javanmard, A. and Montanari, A. (2013), Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. arXiv: 1301.4240.

Javanmard, A. and Montanari, A. (2014), Confidence intervals and hypothe-

sis testing for high-dimensional regression, *Journal of Machine Learning Research* **15**, 2869–2909.

Jerrum, M. and Sinclair, A. (1993), Polynomial-time approximation algorithms for the Ising model, *SIAM Journal of Computing* **22**, 1087–1116.

Jerrum, M. and Sinclair, A. (1996), The Markov chain Monte Carlo method: An approach to approximate counting and integration, in D. Hochbaum, ed., *Approximation algorithms for NP-hard problems*, PWS Publishing, Boston.

Johnson, N. (2013), A dynamic programming algorithm for the fused Lasso and ℓ_0-segmentation, *Journal of Computational and Graphical Statistics* **22**(2), 246–260.

Johnson, W. B. and Lindenstrauss, J. (1984), Extensions of Lipschitz mappings into a Hilbert space, *Contemporary Mathematics* **26**, 189–206.

Johnstone, I. (2001), On the distribution of the largest eigenvalue in principal components analysis, *Annals of Statistics* **29**(2), 295–327.

Johnstone, I. and Lu, A. (2009), On consistency and sparsity for principal components analysis in high dimensions, *Journal of the American Statistical Association* **104**, 682–693.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003), A modified principal component technique based on the Lasso, *Journal of Computational and Graphical Statistics* **12**, 531–547.

Kaiser, H. (1958), The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

Kalisch, M. and Bühlmann, P. (2007), Estimating high-dimensional directed acyclic graphs with the PC algorithm, *Journal of Machine Learning Research* **8**, 613–636.

Kastelyn, P. W. (1963), Dimer statistics and phase transitions, *Journal of Mathematical Physics* **4**, 287–293.

Keshavan, R. H., Montanari, A. and Oh, S. (2010), Matrix completion from noisy entries, *Journal of Machine Learning Research* **11**, 2057–2078.

Keshavan, R. H., Oh, S. and Montanari, A. (2009), Matrix completion from a few entries, *IEEE Transactions on Information Theory* **56(6)**, 2980–2998.

Kim, S., Koh, K., Boyd, S. and Gorinevsky, D. (2009), L1 trend filtering, *SIAM Review, problems and techniques section* **51**(2), 339–360.

Knight, K. and Fu, W. J. (2000), Asymptotics for Lasso-type estimators, *Annals of Statistics* **28**, 1356–1378.

Koh, K., Kim, S. and Boyd, S. (2007), An interior-point method for large-scale ℓ_1-regularized logistic regression, *Journal of Machine Learning Research* **8**, 1519–1555.

Koller, D. and Friedman, N. (2009), *Probabilistic Graphical Models*, The MIT

Press, Cambridge MA.

Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011), Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, *Annals of Statistics* **39**, 2302–2329.

Koltchinskii, V. and Yuan, M. (2008), Sparse recovery in large ensembles of kernel machines, in *Proceedings of the Annual Conference on Learning Theory (COLT)*.

Koltchinskii, V. and Yuan, M. (2010), Sparsity in multiple kernel learning, *Annals of Statistics* **38**, 3660–3695.

Krahmer, F. and Ward, R. (2011), New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property, *SIAM Journal on Mathematical Analysis* **43**(3), 1269–1281.

Lang, K. (1995), Newsweeder: Learning to filter netnews., in *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pp. 331–339.

Lange, K. (2004), *Optimization*, Springer, New York.

Lange, K., Hunter, D. R. and Yang, I. (2000), Optimization transfer using surrogate objective functions (with discussion), *Computational and Graphical Statistics* **9**, 1–59.

Laurent, M. (2001), Matrix completion problems, in *The Encyclopedia of Optimization*, Kluwer Academic, pp. 221–229.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford University Press.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal of the Royal Statistical Society Series B* **50**, 155–224.

Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. and Jackel, L. (1990), Handwritten digit recognition with a back-propogation network, in D. Touretzky, ed., *Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kaufman, Denver, CO, pp. 386–404.

Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J. and Ng, A. (2012), Building high-level features using large scale unsupervised learning, in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland.

Lee, J. and Hastie, T. (2014), Learning the structure of mixed graphical models, *Journal of Computational and Graphical Statistics* . advanced online access.

Lee, J., Sun, D., Sun, Y. and Taylor, J. (2013), Exact post-selection inference, with application to the Lasso. arXiv:1311.6238.

Lee, J., Sun, Y. and Saunders, M. (2014), Proximal newton-type meth-

ods for minimizing composite functions, *SIAM Journal on Optimization* **24**(3), 1420–1443.

Lee, J., Sun, Y. and Taylor, J. (2013), On model selection consistency of m-estimators with geometrically decomposable penalties, Technical report, Stanford University. arxiv1305.7477v4.

Lee, M., Shen, H., Huang, J. and Marron, J. (2010), Biclustering via sparse singular value decomposition, *Biometrics* pp. 1086–1095.

Lee, S., Lee, H., Abneel, P. and Ng, A. (2006), Efficient L1 logistic regression, in *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*.

Lei, J. and Vu, V. Q. (2015), Sparsistency and Agnostic Inference in Sparse PCA, *Ann. Statist.* **43**(1), 299–322.

Leng, C. (2008), Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data, *Computational Biology and Chemistry* **32**, 417–425.

Li, L., Huang, W., Gu, I. Y. and Tian, Q. (2004), Statistical modeling of complex backgrounds for foreground object detection, *IEEE Transactions on Image Processing* **13**(11), 1459–1472.

Lim, M. and Hastie, T. (2014), Learning interactions via hierarchical group-Lasso regularization, *Journal of Computational and Graphical Statistics (online access)* .

Lin, Y. and Zhang, H. H. (2003), Component selection and smoothing in smoothing spline analysis of variance models, Technical report, Department of Statistics, University of Wisconsin, Madison.

Lockhart, R., Taylor, J., Tibshirani$_2$, R. and Tibshirani, R. (2014), A significance test for the Lasso, *Annals of Statistics (with discussion)* **42**(2), 413–468.

Loftus, J. and Taylor, J. (2014), A significance test for forward stepwise model selection. arXiv:1405.3920.

Lounici, K., Pontil, M., Tsybakov, A. and van de Geer, S. (2009), Taking advantage of sparsity in multi-task learning, Technical report, ETH Zurich.

Lustig, M., Donoho, D., Santos, J. and Pauly, J. (2008), Compressed sensing MRI, *IEEE Signal Processing Magazine* **27**, 72–82.

Lykou, A. and Whittaker, J. (2010), Sparse CCA using a Lasso with positivity constraints, *Computational Statistics & Data Analysis* **54**(12), 3144–3157.

Ma, S., Xue, L. and Zou, H. (2013), Alternating direction methods for latent variable Gaussian graphical model selection, *Neural Computation* **25**, 2172–2198.

Ma, Z. (2010), Contributions to high-dimensional principal component analysis, PhD thesis, Department of Statistics, Stanford University.

Ma, Z. (2013), Sparse principal component analysis and iterative thresholding, *Annals of Statistics* **41**(2), 772–801.

Mahoney, M. W. (2011), Randomized algorithms for matrices and data, *Foundations and Trends in Machine Learning in Machine Learning* **3**(2).

Mangasarian, O. (1999), Arbitrary-norm separating plane., *Operations Research Letters* **24**(1-2), 15–23.

Mazumder, R., Friedman, J. and Hastie, T. (2011), *Sparsenet*: Coordinate descent with non-convex penalties, *Journal of the American Statistical Association* **106**(495), 1125–1138.

Mazumder, R. and Hastie, T. (2012), The Graphical Lasso: New insights and alternatives, *Electronic Journal of Statistics* **6**, 2125–2149.

Mazumder, R., Hastie, T. and Friedman, J. (2012), *sparsenet: Fit sparse linear regression models via nonconvex optimization*. R package version 1.0. **URL:** *http://CRAN.R-project.org/package=sparsenet*

Mazumder, R., Hastie, T. and Tibshirani, R. (2010), Spectral regularization algorithms for learning large incomplete matrices, *Journal of Machine Learning Research* **11**, 2287–2322.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman & Hall, London.

Meier, L., van de Geer, S. and Bühlmann, P. (2008), The group Lasso for logistic regression, *Journal of the Royal Statistical Society B* **70**(1), 53–71.

Meier, L., van de Geer, S. and Bühlmann, P. (2009), High-dimensional additive modeling, *Annals of Statistics* **37**, 3779–3821.

Meinshausen, N. (2007), Relaxed Lasso, *Computational Statistics and Data Analysis* pp. 374–393.

Meinshausen, N. and Bühlmann, P. (2006), High-dimensional graphs and variable selection with the Lasso, *Annals of Statistics* **34**, 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010), Stability selection, *Journal of the Royal Statistical Society Series B* **72**(4), 417–473.

Mézard, M. and Montanari, A. (2008), *Information, Physics and Computation*, Oxford University Press, New York, NY.

Negahban, S., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *Statistical Science* **27**(4), 538–557.

Negahban, S. and Wainwright, M. J. (2011*a*), Estimation of (near) low-rank matrices with noise and high-dimensional scaling, *Annals of Statistics* **39**(2), 1069–1097.

Negahban, S. and Wainwright, M. J. (2011*b*), Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization, *IEEE Transactions on Information Theory* **57**(6), 3481–3863.

Negahban, S. and Wainwright, M. J. (2012), Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise, *Journal of Machine Learning Research* **13**, 1665–1697.

Nelder, J. and Wedderburn, R. (1972), Generalized linear models, *J. Royal Statist. Soc. B.* **135**(3), 370–384.

Nemirovski, A. and Yudin, D. B. (1983), *Problem Complexity and Method Efficiency in Optimization*, John Wiley and Sons, New York.

Nesterov, Y. (2004), *Introductory Lectures on Convex Optimization*, Kluwer Academic Publishers, New York.

Nesterov, Y. (2007), Gradient methods for minimizing composite objective function, Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL).

Netrapalli, P., Jain, P. and Sanghavi, S. (2013), Phase retrieval using alternating minimization, in *Advances in Neural Information Processing Systems (NIPS Conference Proceedings)*, pp. 2796–2804.

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011), Union support recovery in high-dimensional multivariate regression, *Annals of Statistics* **39**(1), 1–47.

Oldenburg, D. W., Scheuer, T. and Levy, S. (1983), Recovery of the acoustic impedance from reflection seismograms, *Geophysics* **48**(10), 1318–1337.

Olsen, S. (2002), 'Amazon blushes over sex link gaffe', CNET News. http://news.cnet.com/2100-1023-976435.html.

Olshausen, B. and Field, D. (1996), Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature* **381**.

Park, T. and Casella, G. (2008), The Bayesian Lasso, *Journal of the American Statistical Association* **103**(482), 681–686.

Parkhomenko, E., Tritchler, D. and Beyene, J. (2009), Sparse canonical correlation analysis with application to genomic data integration, *Statistical Applications in Genetics and Molecular Biology* **8**, 1–34.

Paul, D. and Johnstone, I. (2008), Augmented sparse principal component analysis for high-dimensional data, Technical report, UC Davis.

Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Pelckmans, K., De Moor, B. and Suykens, J. (2005), Convex clustering shrinkage, in *Workshop on Statistics and Optimization of Clustering (PASCAL), London, UK.*

Phardoon, D. and Shawe-Taylor, J. (2009), Sparse canonical correlation analysis. arXiv:0908.2724v1.

Pilanci, M. and Wainwright, M. J. (2014), Randomized sketches of convex programs with sharp guarantees, Technical report, UC Berkeley. Full length version at arXiv:1404.7203; Presented in part at ISIT 2014.

Puig, A., Wiesel, A. and Hero, A. (2009), A multidimensional shrinkage thresholding operator, in *Proceedings of the 15th workshop on Statistical Signal Processing, SSP'09*, IEEE, pp. 113–116.

Raskutti, G., Wainwright, M. J. and Yu, B. (2009), Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness, in *Advances in Neural Information Processing Systems 22*, MIT Press, Cambridge MA., pp. 1563–1570.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010), Restricted eigenvalue conditions for correlated Gaussian designs, *Journal of Machine Learning Research* **11**, 2241–2259.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011), Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls, *IEEE Transactions on Information Theory* **57**(10), 6976–6994.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012), Minimax-optimal rates for sparse additive models over kernel classes via convex programming, *Journal of Machine Learning Research* **12**, 389–427.

Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2009), Sparse additive models, *Journal of the Royal Statistical Society Series B.* **71**(5), 1009–1030.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2010), High-dimensional ising model selection using ℓ_1-regularized logistic regression, *Annals of Statistics* **38**(3), 1287–1319.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011), High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence, *Electronic Journal of Statistics* **5**, 935–980.

Recht, B. (2011), A simpler approach to matrix completion, *Journal of Machine Learning Research* **12**, 3413–3430.

Recht, B., Fazel, M. and Parrilo, P. A. (2010), Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, *SIAM Review* **52**(3), 471–501.

Rennie, J. and Srebro, N. (2005), Fast maximum margin matrix factorization for collaborative prediction, in *Proceedings of the 22nd International Conference on Machine Learning*, Association for Computing Machinery, pp. 713–719.

Rish, I. and Grabarnik, G. (2014), *Sparse Modeling: Theory, Algorithms, and Applications*, Chapman and Hall/CRC.

Rockafellar, R. T. (1996), *Convex Analysis*, Princeton University Press.

Rohde, A. and Tsybakov, A. (2011), Estimation of high-dimensional low-rank matrices, *Annals of Statistics* **39**(2), 887–930.

Rosset, S. and Zhu, J. (2007), Adaptable, efficient and robust methods for regression and classification via piecewise linear regularized coefficient

paths, *Annals of Statistics* **35**(3).

Rosset, S., Zhu, J. and Hastie, T. (2004), Boosting as a regularized path to a maximum margin classifier, *Journal of Machine Learning Research* **5**, 941–973.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008), Sparse permutation invariant covariance estimation, *Electronic Journal of Statistics* **2**, 494–515.

Rubin, D. (1981), The Bayesian Bootstrap, *Annals of Statistics* **9**, 130–134.

Rudelson, M. and Zhou, S. (2013), Reconstruction from anisotropic random measurements, *IEEE Transactions on Information Theory* **59**(6), 3434–3447.

Ruderman, D. (1994), The statistics of natural images, *Network: Computation in Neural Systems* **5**, 517–548.

Santhanam, N. P. and Wainwright, M. J. (2008), Information-theoretic limits of high-dimensional model selection, in *International Symposium on Information Theory*, Toronto, Canada.

Santosa, F. and Symes, W. W. (1986), Linear inversion of band-limited reflection seismograms, *SIAM Journal of Scientific and Statistical Computing* **7**(4), 1307–1330.

Scheffé, H. (1953), A method for judging all contrasts in the analysis of variance, *Biometrika* **40**, 87–104.

Schmidt, M., Niculescu-Mizil, A. and Murphy, K. (2007), Learning graphical model structure using l1-regularization paths, in *AAAI proceedings*.
URL: *http://www.cs.ubc.ca/ murphyk/Papers/aaai07.pdf*

Shalev-Shwartz, S., Singer, Y. and Srebro, N. (2007), Pegasos: Primal estimated sub-gradient solver for SVM, in *Proceedings of the 24th international conference on Machine learning*, pp. 807–814.

She, Y. and Owen, A. B. (2011), Outlier detection using nonconvex penalized regression, *Journal of the American Statistical Association* **106**(494), 626–639.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011), Regularization paths for Cox's proportional hazards model via coordinate descent, *Journal of Statistical Software* **39**(5), 1–13.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013), A sparse-group Lasso, *Journal of Computational and Graphical Statistics* **22**(2), 231–245.

Simon, N. and Tibshirani, R. (2012), Standardization and the group Lasso penalty, *Statistica Sinica* **22**, 983–1001.

Simoncelli, E. P. (2005), Statistical modeling of photographic images, in *Handbook of Video and Image Processing, 2nd Edition*, Academic Press, Waltham MA, pp. 431–441.

Simoncelli, E. P. and Freeman, W. T. (1995), The steerable pyramid: A flexible

architecture for multi-scale derivative computation, in *Int'l Conference on Image Processing*, Vol. III, IEEE Sig Proc Soc., Washington, DC, pp. 444–447.

Singer, Y. and Dubiner, M. (2011), Entire relaxation path for maximum entropy models, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNPL 2011)*, pp. 941–948.

Srebro, N., Alon, N. and Jaakkola, T. (2005), Generalization error bounds for collaborative prediction with low-rank matrices, *Advances in Neural Information Processing Systems* .

Srebro, N. and Jaakkola, T. (2003), Weighted low-rank approximations, in *Twentieth International Conference on Machine Learning*, AAAI Press, pp. 720–727.

Srebro, N., Rennie, J. and Jaakkola, T. (2005), Maximum margin matrix factorization, *Advances in Neural Information Processing Systems* **17**, 1329–1336.

Stein, C. (1981), Estimation of the mean of a multivariate normal distribution, *Annals of Statistics* **9**, 1131–1151.

Stone, C. J. (1985), Additive regression and other non-parametric models, *Annals of Statistics* **13**(2), 689–705.

Taylor, J., Lockhart, R., Tibshirani$_2$, R. and Tibshirani, R. (2014), Post-selection adaptive inference for least angle regression and the Lasso. arXiv: 1401.3889; submitted.

Taylor, J., Loftus, J. and Tibshirani$_2$, R. (2013), Tests in adaptive regression via the Kac-Rice formula. arXiv:1308.3020; submitted.

Thodberg, H. H. and Olafsdottir, H. (2003), Adding curvature to minimum description length shape models, in *British Machine Vision Conference (BMVC)*, pp. 251–260.

Thomas, G. S. (1990), *The Rating Guide to Life in America's Small Cities*, Prometheus books. `http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html`.

Tibshirani, R. (1996), Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani$_2$, R. (2012), Strong rules for discarding predictors in Lasso-type problems, *Journal of the Royal Statistical Society Series B*. pp. 245–266.

Tibshirani, R. and Efron, B. (2002), Pre-validation and inference in microarrays, *Statistical Applications in Genetics and Molecular Biology* pp. 1–15.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2001), Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceed-*

ings of the National Academy of Sciences **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003), Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science* pp. 104–117.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005), Sparsity and smoothness via the fused Lasso, *Journal of the Royal Statistical Society, Series B* **67**, 91–108.

Tibshirani$_2$, R. (2013), The Lasso problem and uniqueness, *Electronic Journal of Statistics* **7**, 1456–1490.

Tibshirani$_2$, R. (2014), Adaptive piecewise polynomial estimation via trend filtering, *Annals of Statistics* **42**(1), 285–323.

Tibshirani$_2$, R., Hoefling, H. and Tibshirani, R. (2011), Nearly-isotonic regression, *Technometrics* **53**(1), 54–61.

Tibshirani$_2$, R. and Taylor, J. (2011), The solution path of the generalized Lasso, *Annals of Statistics* **39**(3), 1335–1371.

Tibshirani$_2$, R. and Taylor, J. (2012), Degrees of freedom in Lasso problems, *Annals of Statistics* **40**(2), 1198–1232.

Trendafilov, N. T. and Jolliffe, I. T. (2007), DALASS: Variable selection in discriminant analysis via the LASSO, *Computational Statistics and Data Analysis* **51**, 3718–3736.

Tropp, J. A. (2006), Just relax: Convex programming methods for identifying sparse signals in noise, *IEEE Transactions on Information Theory* **52**(3), 1030–1051.

Tseng, P. (1988), Coordinate ascent for maximizing nondifferentiable concave functions, Technical Report LIDS-P ; 1840, Massachusetts Institute of Technology. Laboratory for Information and Decision Systems.

Tseng, P. (1993), Dual coordinate ascent methods for non-strictly convex minimization, *Mathematical Programming* **59**, 231–247.

Tseng, P. (2001), Convergence of block coordinate descent method for nondifferentiable maximization, *Journal of Optimization Theory and Applications* **109**(3), 474–494.

van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge University Press.

van de Geer, S. and Bühlmann, P. (2009), On the conditions used to prove oracle results for the Lasso, *Electronic Journal of Statistics* **3**, 1360–1392.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2013), On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv: 1303.0518v2.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J. and Wessels, L. F. A. (2006), Cross-validated Cox regression on microarray gene expression data, *Statistics in Medicine* **45**, 3201–3216.

Vandenberghe, L., Boyd, S. and Wu, S. (1998), Determinant maximization with linear matrix inequality constraints, *SIAM Journal on Matrix Analysis and Applications* **19**, 499–533.

Vapnik, V. (1996), *The Nature of Statistical Learning Theory*, Springer, New York.

Vempala, S. (2004), *The Random Projection Method*, Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Providence, RI.

Vershynin, R. (2012), Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing: Theory and Applications*, Cambridge University Press.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013), Fantope projection and selection: A near-optimal convex relaxation of sparse PCA, in *Advances in Neural Information Processing Systems (NIPS Conference Proceedings)*, pp. 2670–2678.

Vu, V. Q. and Lei, J. (2012), Minimax rates of estimation for sparse PCA in high dimensions, in *15th Annual Conference on Artificial Intelligence and Statistics*, La Palma, Canary Islands.

Waaijenborg, S., Versélewel de Witt Hamer, P. and Zwinderman, A. (2008), Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis, *Statistical Applications in Genetics and Molecular Biology* **7**, Article 3.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia, PA.

Wainwright, M. J. (2009), Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso), *IEEE Transactions on Information Theory* pp. 2183–2202.

Wainwright, M. J. and Jordan, M. I. (2008), Graphical models, exponential families and variational inference, *Foundations and Trends in Machine Learning* **1**(1–2), 1–305.

Wainwright, M. J., Simoncelli, E. P. and Willsky, A. S. (2001), Random cascades on wavelet trees and their use in modeling and analyzing natural images, *Applied Computational and Harmonic Analysis* **11**, 89–123.

Wang, H. (2014), Coordinate descent algorithm for covariance graphical Lasso, *Statistics and Computing* **24**(4), 521–529.

Wang, J., Lin, B., Gong, P., Wonka, P. and Ye, J. (2013), Lasso screening rules via dual polytope projection, in *Advances in Neural Information Processing Systems (NIPS Conference Proceedings)*, pp. 1070–1078.

Wang, L., Zhu, J. and Zou, H. (2006), The doubly regularized support vector machine, *Statistica Sinica* **16**(2), 589.

Wang, W., Liang, Y. and Xing, E. P. (2013), Block regularized Lasso for

multivariate multiresponse linear regression, in *Proceedings of the 16th International Conference on Artifical Intelligence and Statistics*, Scottsdale, AZ.

Welsh, D. J. A. (1993), *Complexity: Knots, Colourings, and Counting*, LMS Lecture Note Series, Cambridge University Press, Cambridge.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.

Winkler, G. (1995), *Image Analysis, Random Fields, and Dynamic Monte Carlo methods*, Springer-Verlag, New York, NY.

Witten, D. (2011), *penalizedLDA: Penalized classification using Fisher's linear discriminant*. R package version 1.0.
URL: *http://CRAN.R-project.org/package=penalizedLDA*

Witten, D., Friedman, J. and Simon, N. (2011), New insights and faster computations for the graphical Lasso, *Journal of Computational and Graphical Statistics* **20**, 892–200.

Witten, D. and Tibshirani, R. (2009), Extensions of sparse canonical correlation analysis, with application to genomic data, *Statistical Applications in Genetics and Molecular Biology* **8(1)**, Article 28.

Witten, D. and Tibshirani, R. (2010), A framework for feature selection in clustering, *Journal of the American Statistical Association* **105(490)**, 713–726.

Witten, D. and Tibshirani, R. (2011), Penalized classification using Fisher's linear discriminant, *Journal of the Royal Statistical Society Series B* **73**(5), 753–772.

Witten, D., Tibshirani, R. and Hastie, T. (2009), A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biometrika* **10**, 515–534.

Wu, T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009), Genomewide association analysis by Lasso penalized logistic regression, *Bioinformatics* **25**(6), 714–721.

Wu, T. and Lange, K. (2007), The MM alternative to EM. unpublished.

Wu, T. and Lange, K. (2008), Coordinate descent procedures for Lasso penalized regression, *Annals of Applied Statistics* **2**(1), 224–244.

Xu, H., Caramanis, C. and Mannor, S. (2010), Robust regression and Lasso, *IEEE Transactions on Information Theory* **56**(7), 3561–3574.

Xu, H., Caramanis, C. and Sanghavi, S. (2012), Robust PCA via outlier pursuit, *IEEE Transactions on Information Theory* **58**(5), 3047–3064.

Yi, X., Caramanis, C. and Sanghavi, S. (2014), Alternating minimization for mixed linear regression, in *Proceedings of The 31st International Conference on Machine Learning*, pp. 613–621.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007), Dimension reduction

and coefficient estimation in multivariate linear regression, *Journal of the Royal Statistical Society Series B* **69**(3), 329–346.

Yuan, M. and Lin, Y. (2006*a*), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B* **68**(1), 49–67.

Yuan, M. and Lin, Y. (2006*b*), Model selection and estimation in the Gaussian graphical model, *Biometrika* **94**(1), 19–35.

Yuan, M. and Lin, Y. (2006*c*), On the non-negative garrotte estimator, *Journal of the Royal Statistical Society, Series B* **69**(2), 143–161.

Yuan, X. T. and Zhang, T. (2013), Truncated power method for sparse eigenvalue problems, *Journal of Machine Learning Research* **14**, 899–925.

Zhang, C.-H. (2010), Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics* **38**(2), 894–942.

Zhang, C.-H. and Zhang, S. (2014), Confidence intervals for low-dimensional parameters with high-dimensional data, *Journal of the Royal Statistical Society Series B* **76**(1), 217–242.

Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014), Lower bounds on the performance of polynomial-time algorithms for sparse linear regression, in *Proceedings of the Annual Conference on Learning Theory (COLT)*, Barcelona, Spain. Full length version at http://arxiv.org/abs/1402.1918.

Zhao, P., Rocha, G. and Yu, B. (2009), Grouped and hierarchical model selection through composite absolute penalties, *Annals of Statistics* **37**(6A), 3468–3497.

Zhao, P. and Yu, B. (2006), On model selection consistency of Lasso, *Journal of Machine Learning Research* **7**, 2541–2567.

Zhao, Y., Levina, E. and Zhu, J. (2011), Community extraction for social networks, *Proceedings of the National Academy of Sciences* **108**(18), 7321–7326.

Zhou, S., Lafferty, J. and Wasserman, L. (2008), Time-varying undirected graphs, in *Proceedings of the Annual Conference on Learning Theory (COLT)*, Helsinki, Finland.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2004), 1-norm support vector machines, in *Advances in Neural Information Processing Systems*, Vol. 16, pp. 49–56.

Zou, H. (2006), The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B.* **67**(2), 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2006), Sparse principal component analysis, *Journal of Computational and Graphical Statistics* **15**(2), 265–

286.

Zou, H., Hastie, T. and Tibshirani, R. (2007), On the degrees of freedom of the Lasso, *Annals of Statistics* **35**(5), 2173–2192.

Zou, H. and Li, R. (2008), One-step sparse estimates in nonconcave penalized likelihood models, *The Annals of Statistics* **36**(4), 1509–1533.



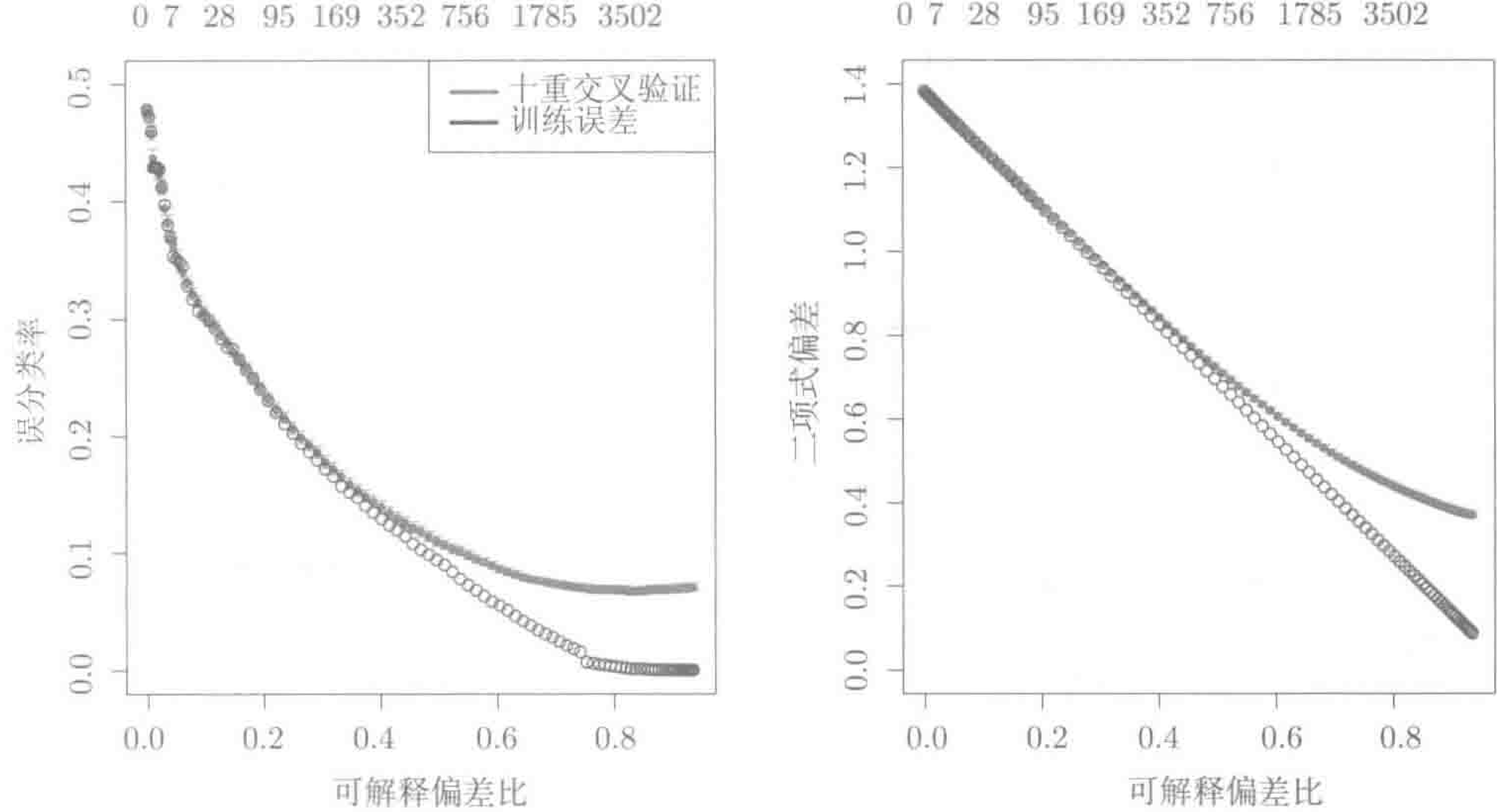


图 3-2　基于 lasso（ℓ_1）惩罚的逻辑斯蒂回归。红线表示新闻组数据的十重交叉验证结果，以及每一点的标准差（在图中并不明显）。左图为误分类率，右图为偏差。图中蓝线为相应的训练误差。图中上侧数字为每个模型的非零参数个数

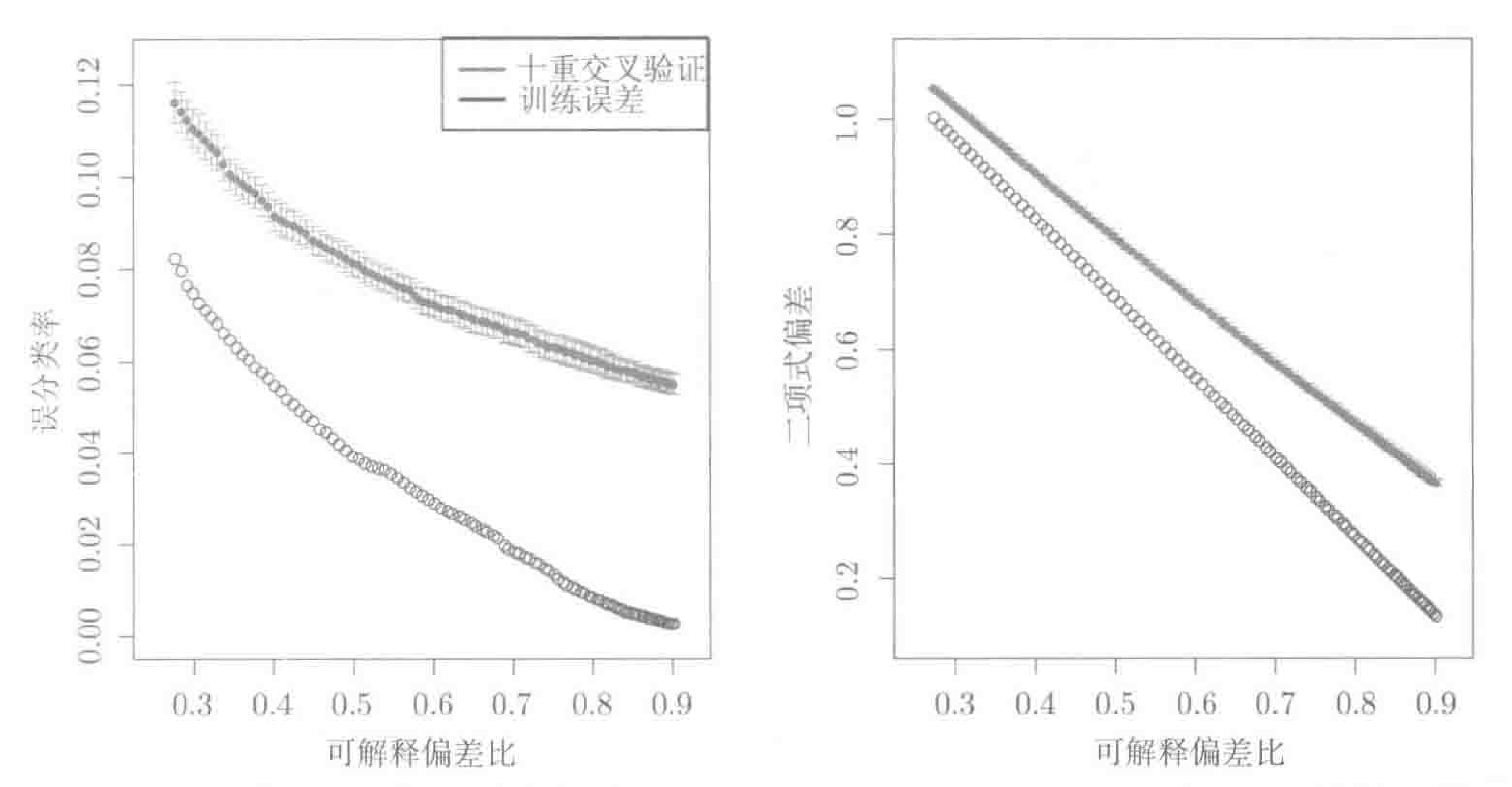


图 3-3　基于岭（ℓ_2）惩罚的逻辑斯蒂回归：红线表示新闻组数据的十重交叉验证结果，以及每一点的标准差范围。左图为误分类率，右图为偏差。图中蓝线为相应的训练误差

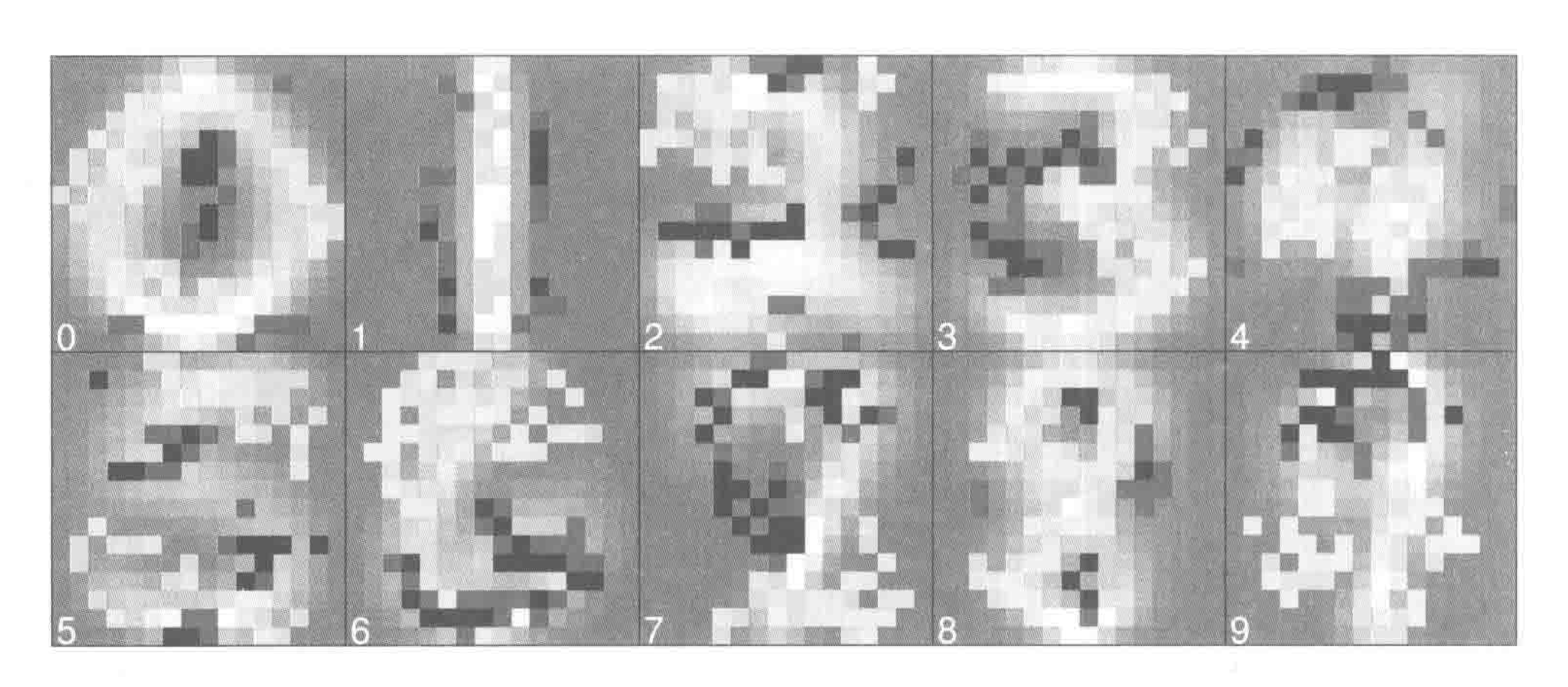


图 3-5　多类 lasso 回归中每一类数字的系数。灰色背景为每一类的平均训练样本。叠加在上的颜色（黄色为正系数，蓝色为负系数）表示为每一类的非零系数。注意，这些值为非零的地方有所不同，由此便产生了每一类的判别评价（score）。并非所有这些值都可解释

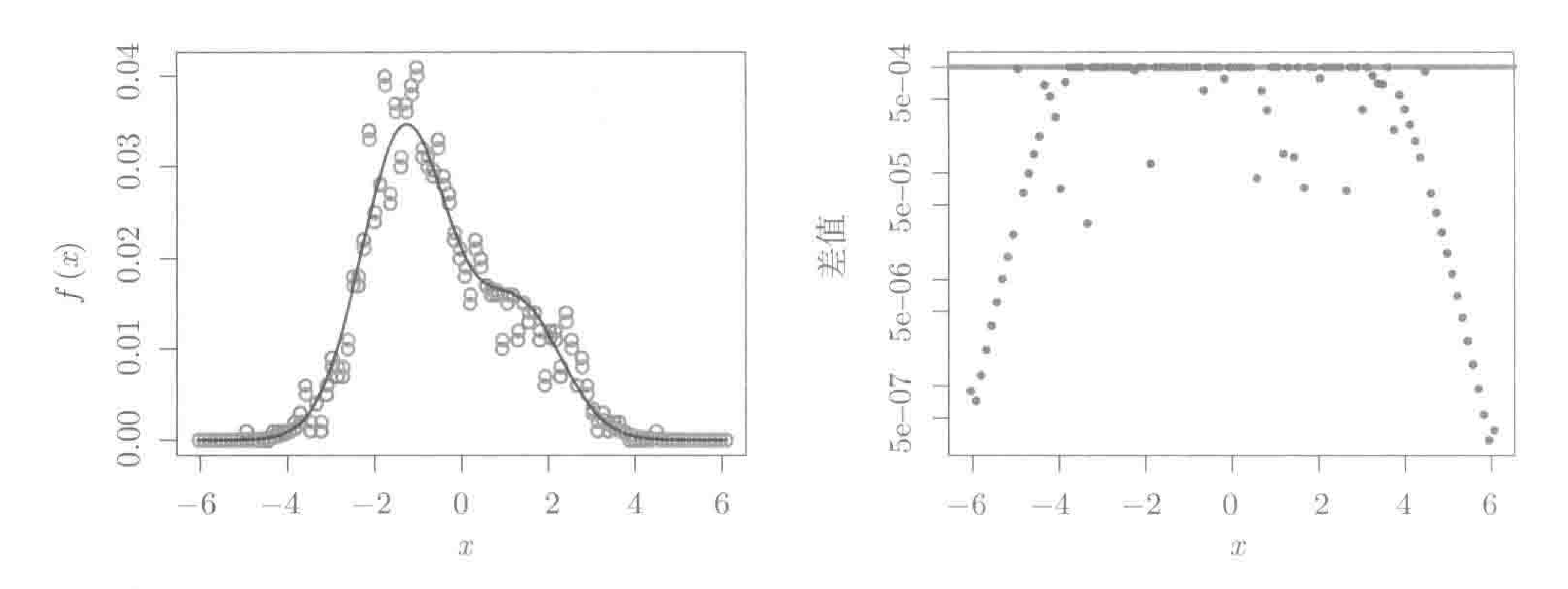


图 3-6　通过泊松模型估计分布。在左图中，黑实线是分布 $\boldsymbol{u}$，这里表示离散化一维分布 $f(x)$ 到 100 个单元。蓝点表示观测到的分布，黄点表示模型恢复后的分布。观测到的分布可能会有一些为零的计数，经过模型恢复后的分布与 $\boldsymbol{u}$ 支撑相同。右图为 $N=100$ 时，$|\widehat{q}_k - r_k|$ 的差，其约束小于 $\delta = 0.001$，即水平黄线以下

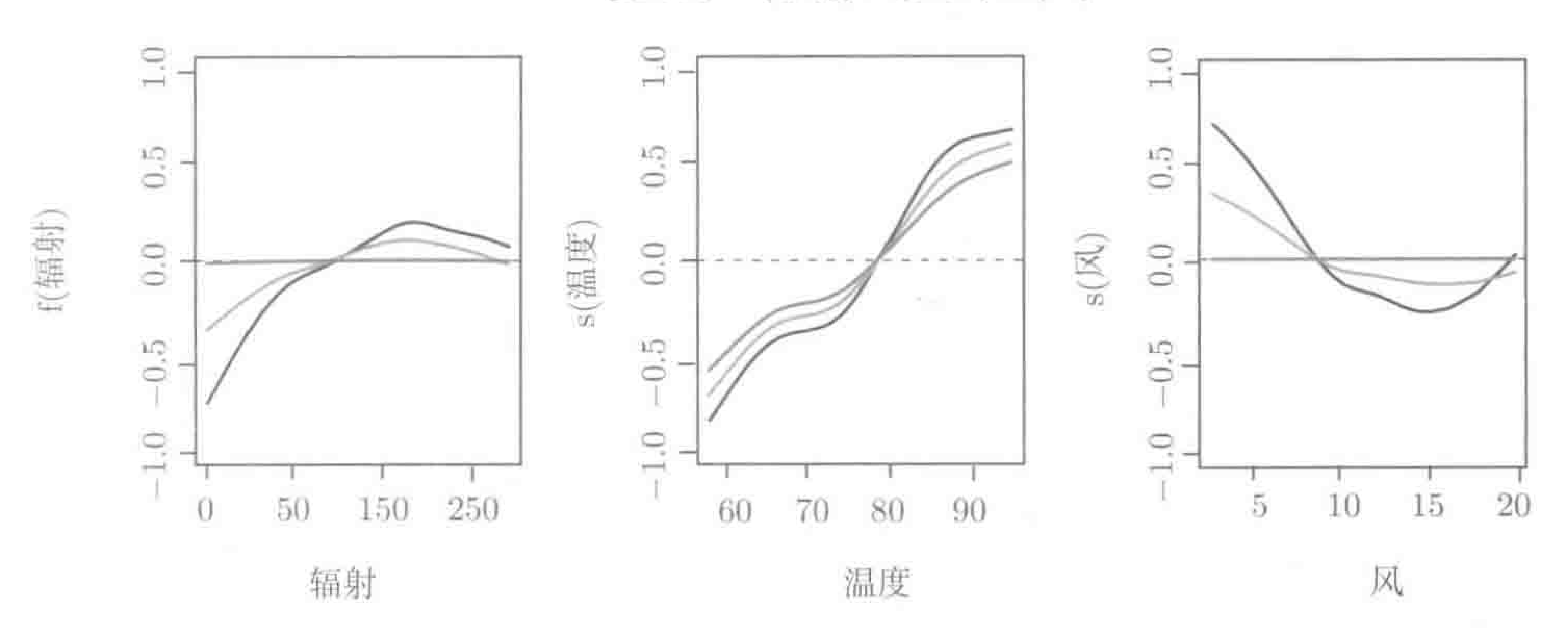


图 4-7 对空气污染数据，采用三个 SPAM 模型进行拟合所得的结果。响应变量是臭氧浓度取对数，有三个预测子：辐射、温度和风速。在拟合加法模型中用到了光滑样条，其 $df=5$。图中三条曲线分别对应 λ=0（黑线），λ=2（橙线），和 λ=4（红线）。可以看到，收缩使得温度函数相对不受影响，辐射和风速方面则影响显著

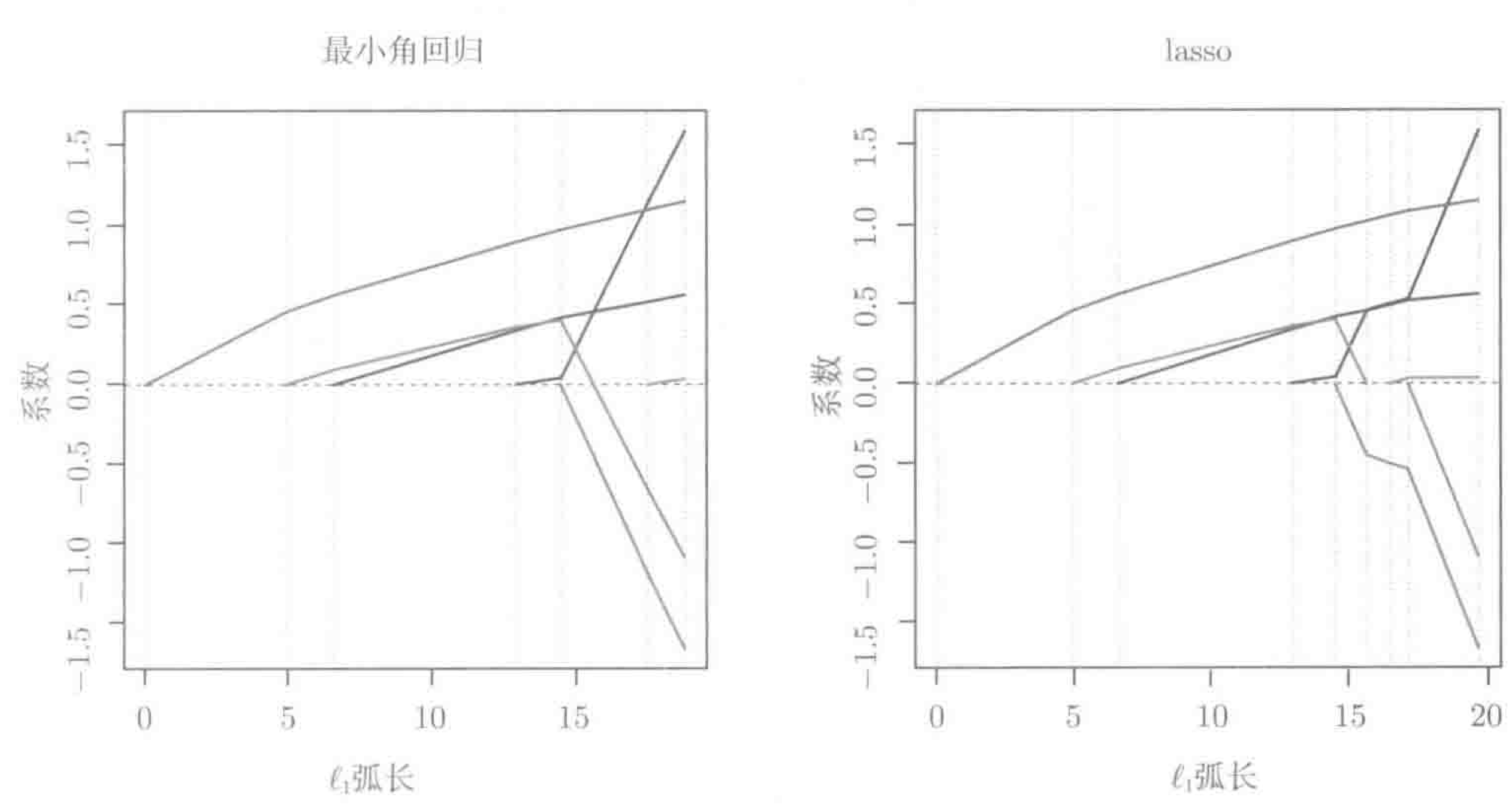


图 5-9 左图为 LAR 在模拟数据上得到的系数轮廓，这些系数与 L^1 弧长之间有函数关系。右图是求解 lasso 问题所得到的系数轮廓。二者在红色系数轮廓穿过水平轴（弧长为 16 附近）之前是一样的

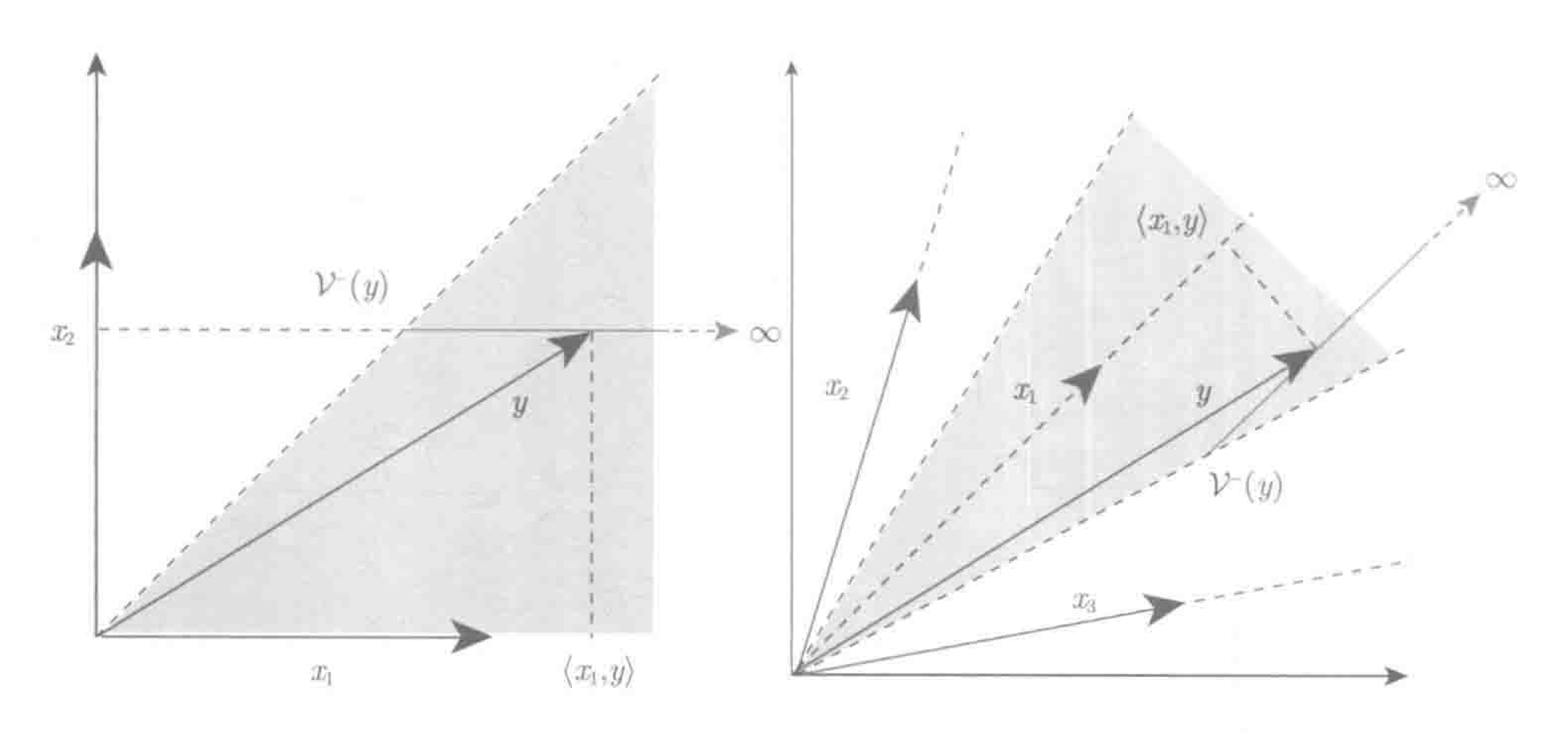


图 6-10　在 $\lambda_1 = \langle \boldsymbol{x}_1, \boldsymbol{y}\rangle$ 时，例 6.1 的选择区域。左图所示为两个正交预测子，右图所示为三个相关预测子。红线表示集合 $P_{\boldsymbol{\eta}^\perp}\boldsymbol{y} + t\boldsymbol{\eta}$ 在选择区域内的部分。左图中 $\nu^-(\boldsymbol{y}) = \langle \boldsymbol{x}_2, \boldsymbol{y}\rangle$，而右图中 $\nu^-(\boldsymbol{y}) = \lambda_2$

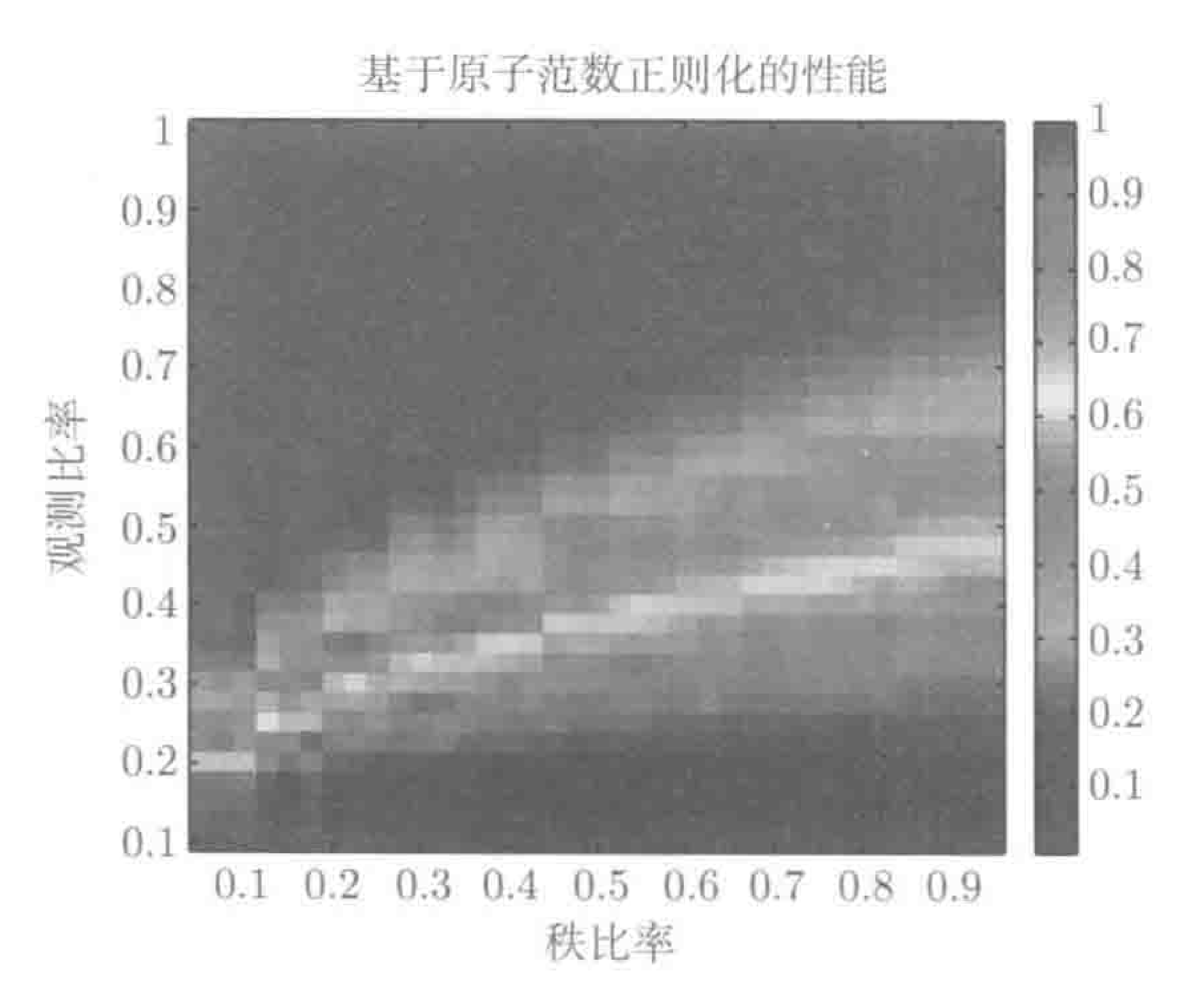


图 7-5　通过 Soft-Impute 算法求解基于原子范数正则化估计子（7.10）的性能。这是一个基于式（7.7）的噪声矩阵填充模型，其中 $\boldsymbol{L}^* = \boldsymbol{C}\boldsymbol{G}^{\mathrm{T}}$ 的秩为 r。这幅图为秩比率 $\delta = \frac{r\log p}{p}$ 和观测比率 $\nu = \frac{N}{p^2}$（该比率表示在 $p \times p$ 矩阵中，观测到的元素所占的比例）的函数曲线图，其中 $p = 50$，采用相对 Frobenius 范数误差 $\frac{\|\hat{\boldsymbol{L}}-\boldsymbol{L}^*\|_F^2}{\|\boldsymbol{L}^*\|_F^2}$ 进行度量。式（7.7）是观测结果的线性形式，其中 $w_{ij} \sim N(0,\sigma^2)$，$\sigma = 1/4$，采用 Soft-Impute 算法求解式（7.10），其中 $\lambda/N = 2\sigma\sqrt{\frac{P}{N}}$，理论上建议选择后者。该理论还指出，只要 $\nu \succ \delta$，Frobenius 误差就会变小，这幅图证实了这个结论

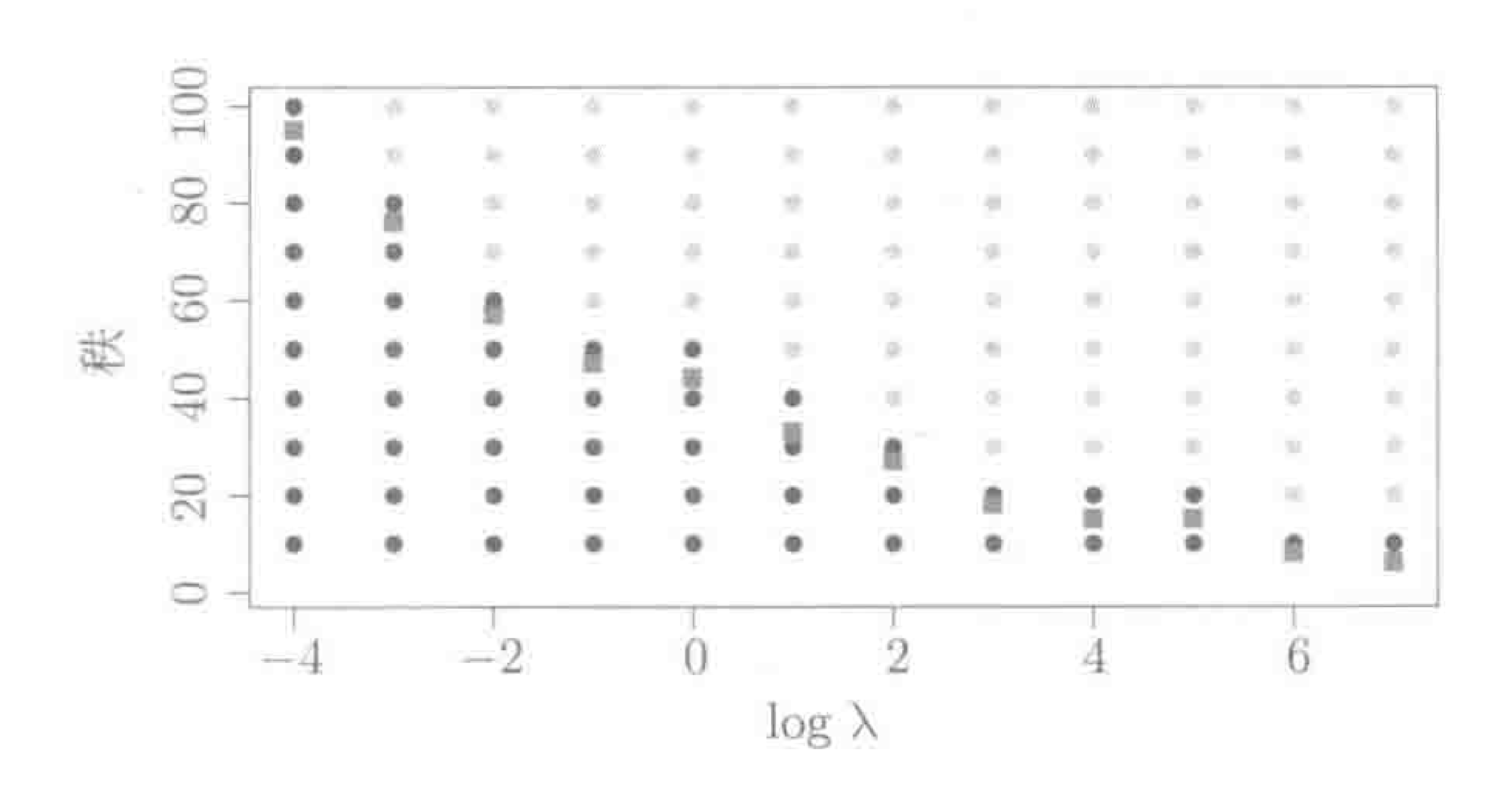


图 7-7　针对一个简单的例子比较 MMMF（灰色和黑色的点）和 Soft-Impute（红色的点）。对于红点上方的秩，MMMF 的解与 Soft-Impute 相同，因此灰点显出了冗余。另一方面，若对 MMMF 固定秩（λ 已指定），且这个秩要比 Soft-Impute 的解小，就会得到一个非凸问题

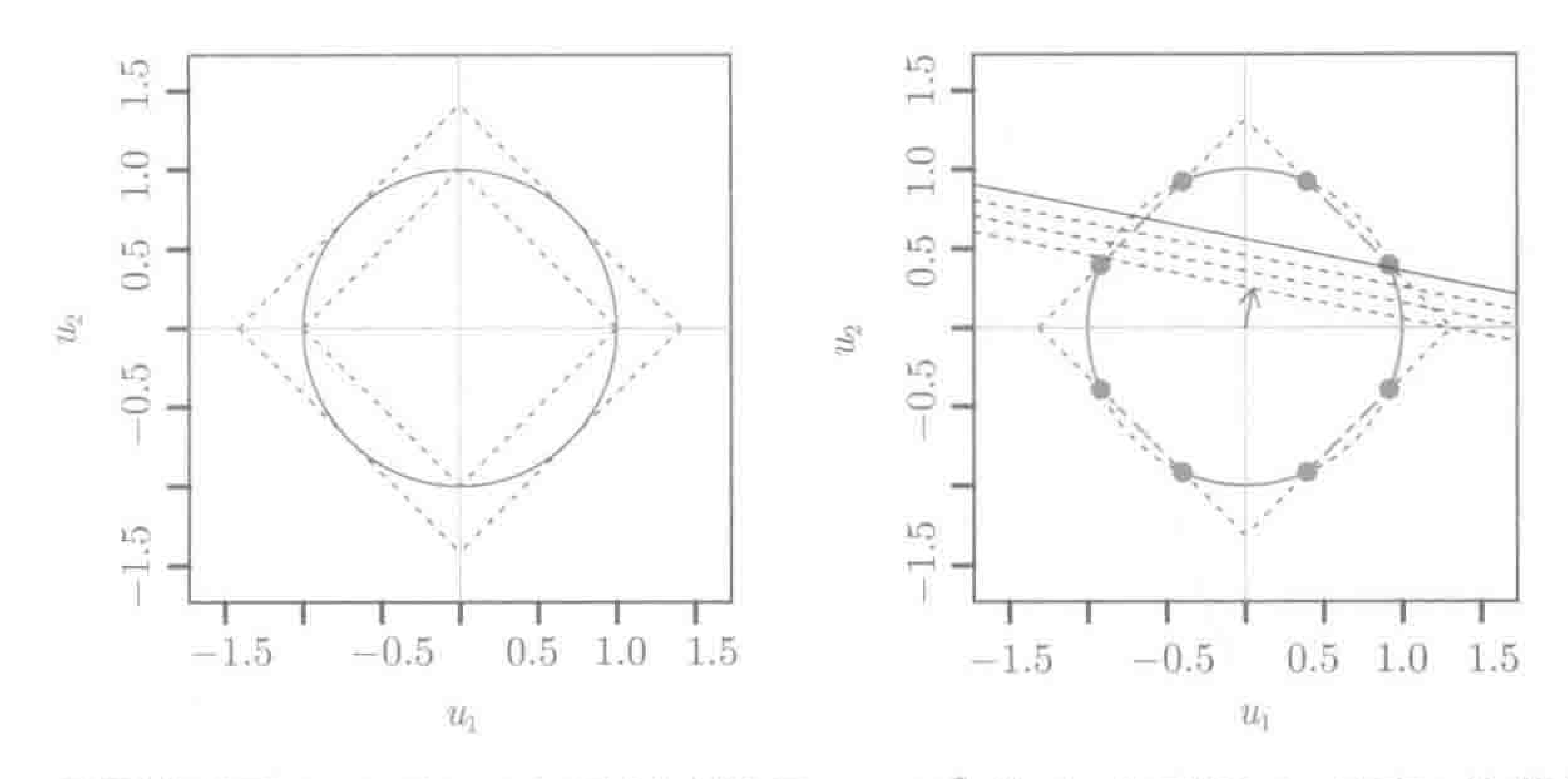


图 7-10　用图形表示 PMD(ℓ_1, ℓ_2) 目标函数下 $\boldsymbol{u} \in \mathbb{R}^2$ 的 ℓ_1 约束和 ℓ_2 约束。这些约束分别为 $\|\boldsymbol{u}\|_2^2 \leqslant 1$ 和 $\|\boldsymbol{u}\|_1 \leqslant c$。一横一纵两条直构成的十字表示坐标轴 u_1 和 u_2。左图的实心圆是 ℓ_2 约束。约束半径必须为 $1 \sim \sqrt{2}$，ℓ_1 约束和 ℓ_2 约束才都会起作用。约束 $\|\boldsymbol{u}\|_1 = 1$ 和 $\|\boldsymbol{u}\|_1 = \sqrt{2}$ 用虚线显示。右图显示 ℓ_1 约束和 ℓ_2 约束，其中 c 为 $1 \sim \sqrt{2}$ 的某个值。红色轮廓为约束区域的边界。黑线是式（7.30）作为 $\boldsymbol{u}$ 的函数的线性轮廓，可将其看成在朝右上方移动。红色实线所构成的弧表示在算法 7.2 中 $\lambda_1 = 0$ 时的解（ℓ_2 约束起作用，而 ℓ_1 约束没有起作用）。该图展示的是二维情形，对于 ℓ_1 约束和 ℓ_2 约束都会起作用的点，它们的坐标 u_1 和 u_2 都不会为 0。没有 ℓ_2 约束，结束总会在拐角处，这样会得到平凡解

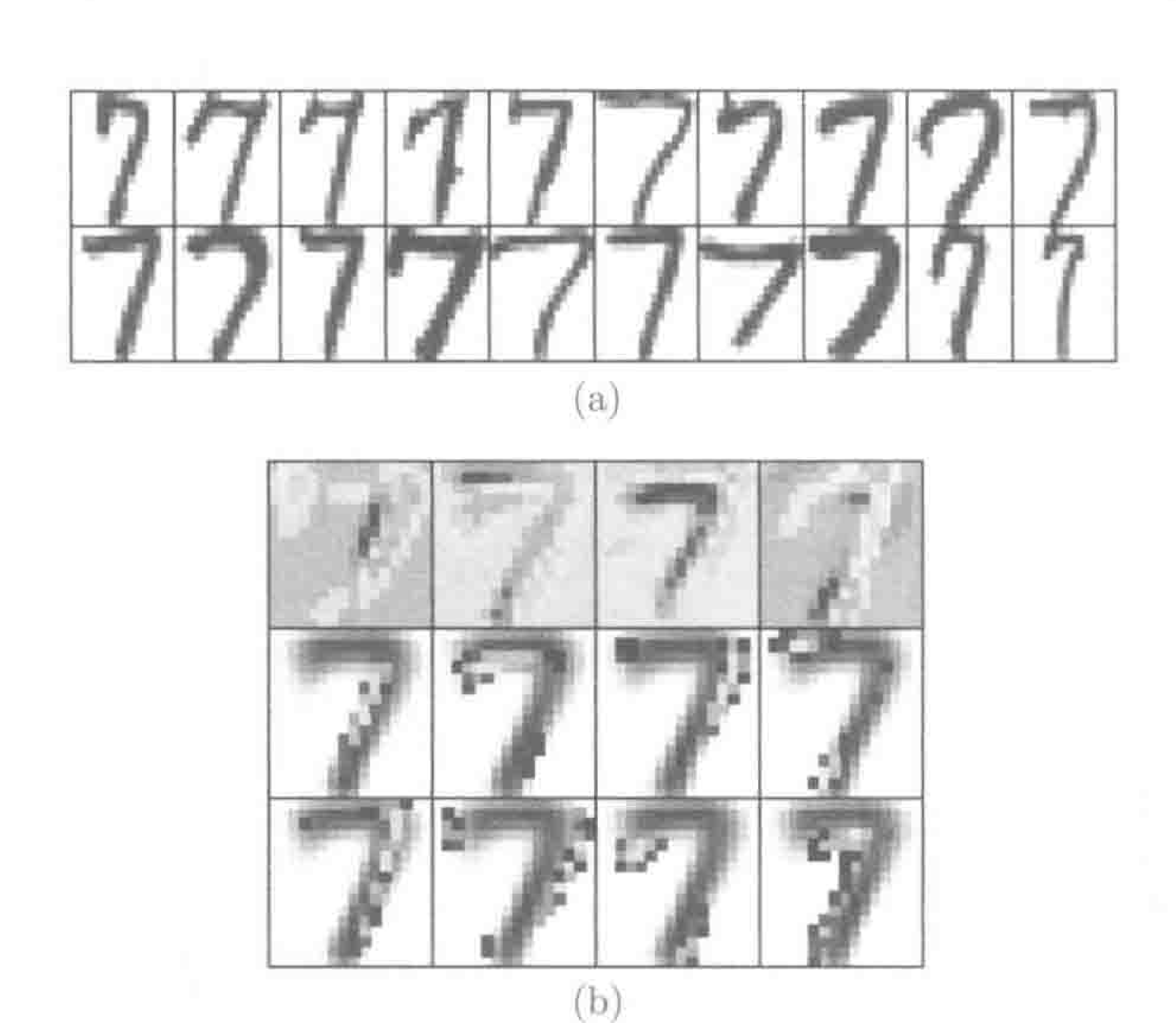


图 8-2　(a) 取自邮政编码数据库中的手写数字 7。(b) 上面一行：手写数字 7 的前 4 个主成分（不同颜色表示不同的载荷：负载荷为黄色；正载荷为蓝色）；下面两行：前 8 个稀疏主成分，载荷被限制为正数。这些稀疏主成分叠加到普通的数字 7 上，以提高解释性

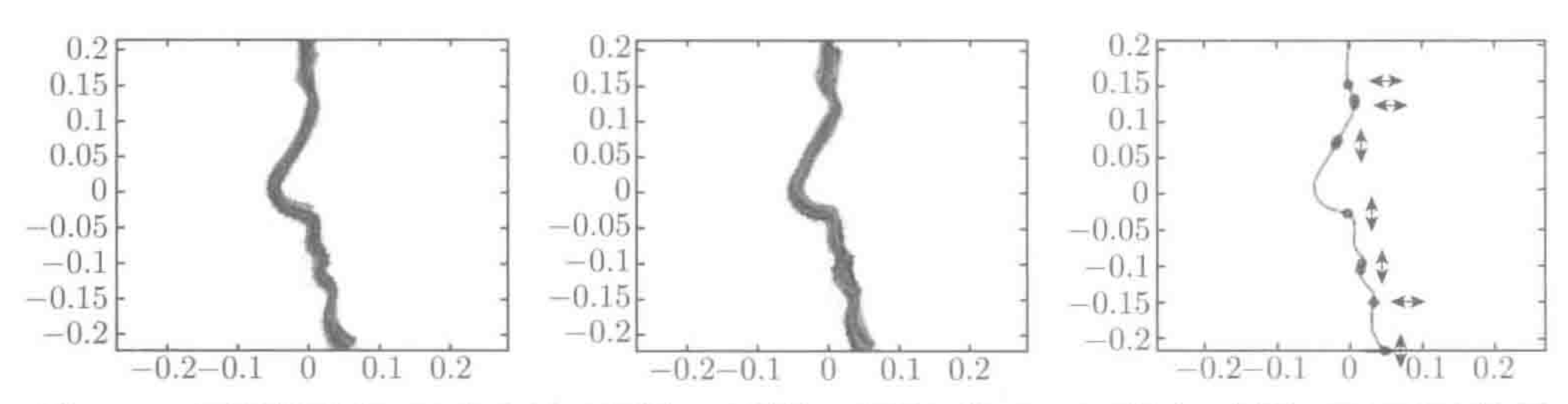


图 8-8　面部轮廓和 65 名女性（左图）及男性（中图）的 (x, y) 坐标。右图：面部轮廓的平均形状，SDA 模型中有 10 个坐标。叠加圆点表示保留在稀疏判别向量中的关键点。箭头表示男性和女性之间不同的方向

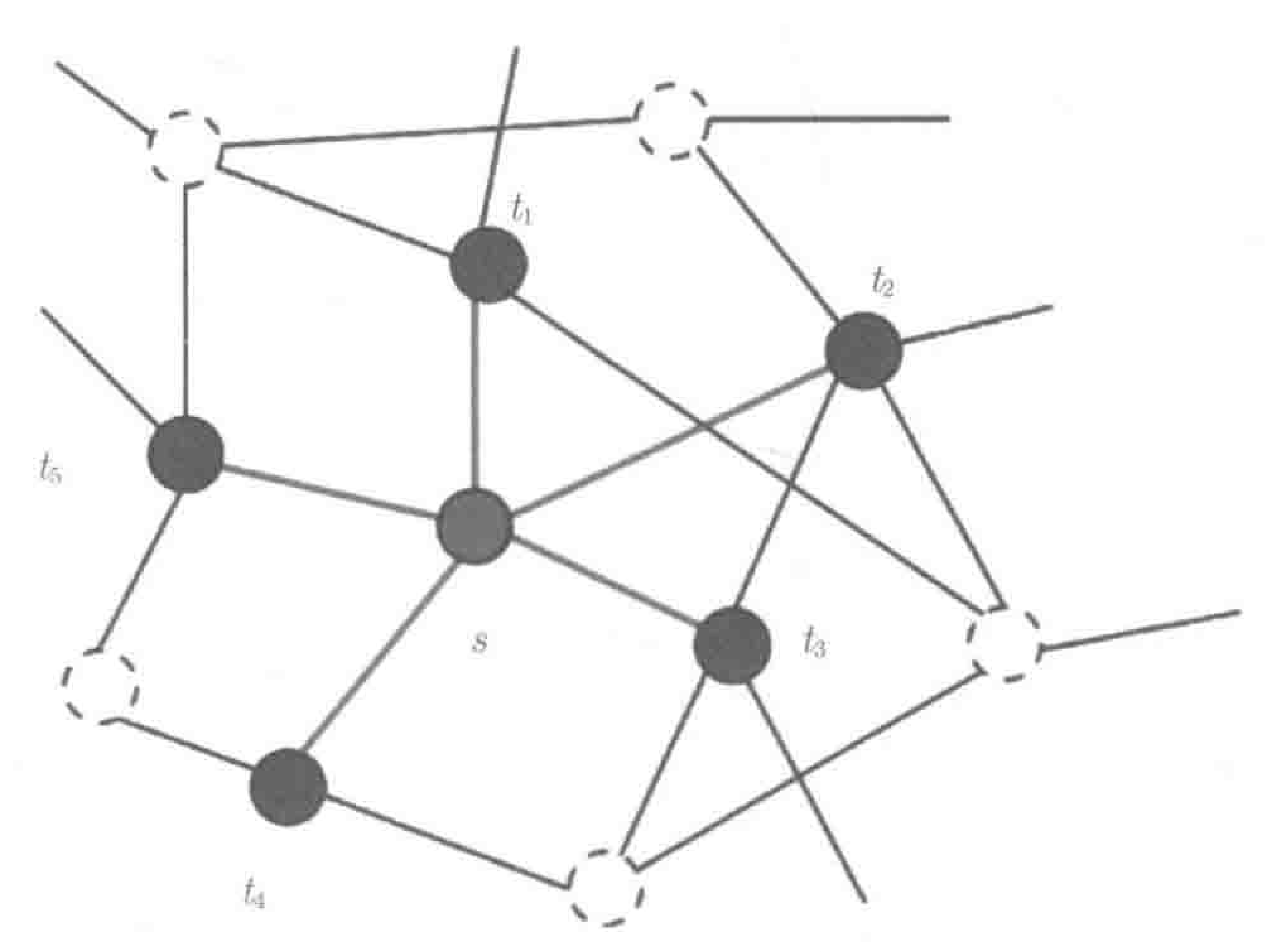


图 9-6　深蓝色顶点形成了灰色顶点的相邻集 $\mathcal{N}(s)$；集合 $\mathcal{N}^{+}(s)$ 由 $\mathcal{N}(s)\cup\{s\}$ 联合而成。注意，$\mathcal{N}(s)$ 是图中的割集，分割 $\{s\}$ 和 $V\backslash\mathcal{N}^{+}(s)$。因此，变量 X_s 在给定相邻集中变量 $X_{\mathcal{N}(s)}$ 下条件独立于 $X_{V\backslash\mathcal{N}^{+}(s)}$。这种条件独立性意味着基于图中其他所有变量的 X_s 的最优预测只依赖于 $X_{\mathcal{N}(s)}$

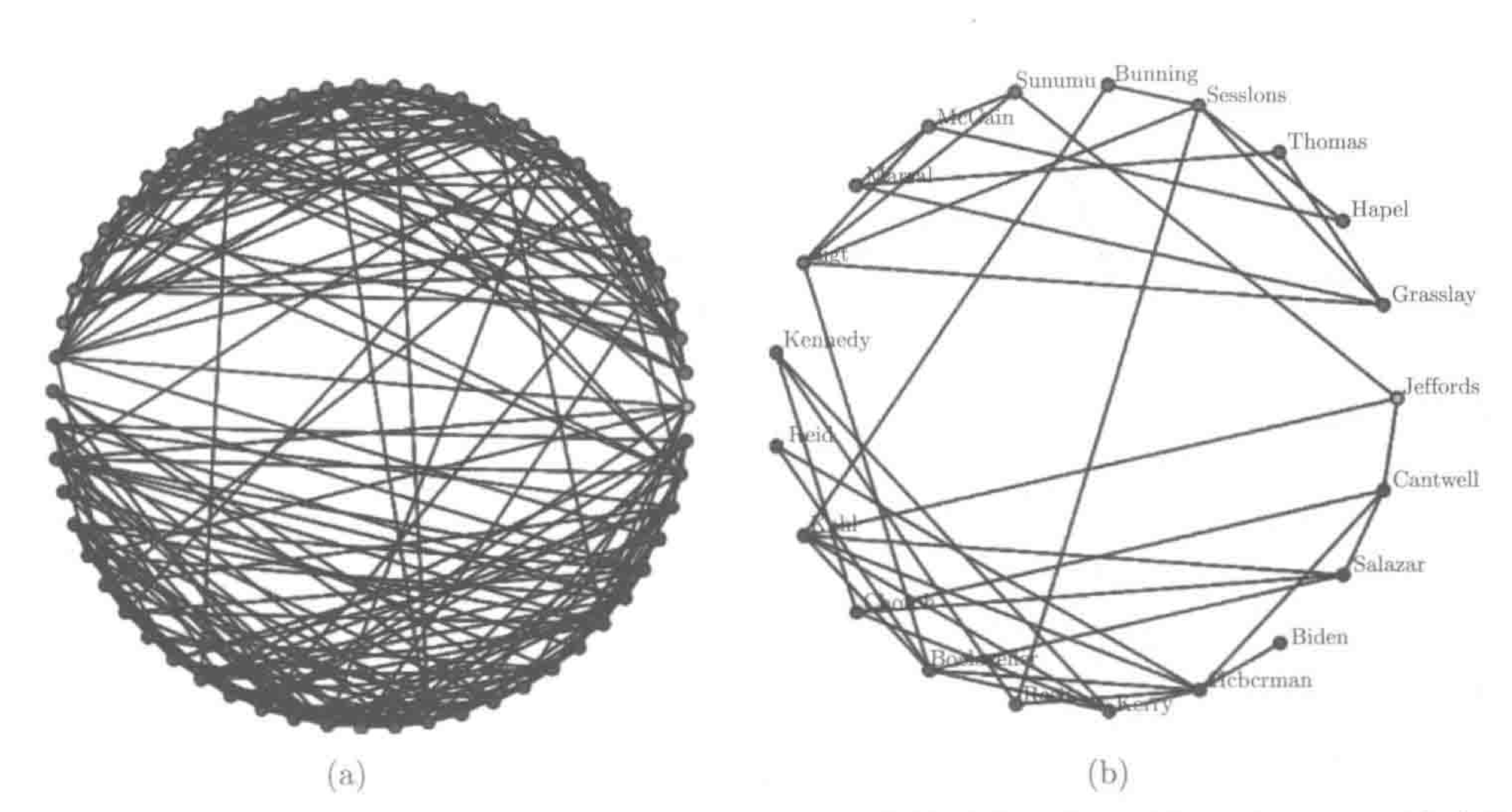


图 9-7　美国参议院（2004—2006）投票数据估计出来的政客网络。数据集为 p=100 个参议员，总共 N=546 场投票，$X_s=+1$（$X_s=-1$）意味着参议员 s 投了“赞成”（“反对”）。这里用基于近邻的逻辑斯蒂回归方法拟合数据得到一对图模型。(a) 拟合包含 55 个参议员的子图，蓝色/红色/黄色分别表示民主/共和/独立党派参议员。注意，子图显示集群根据党派有一个鲜明的两部分趋势。少量的参议员有跨党派的联系。(b) 相同社交网络的更小子图

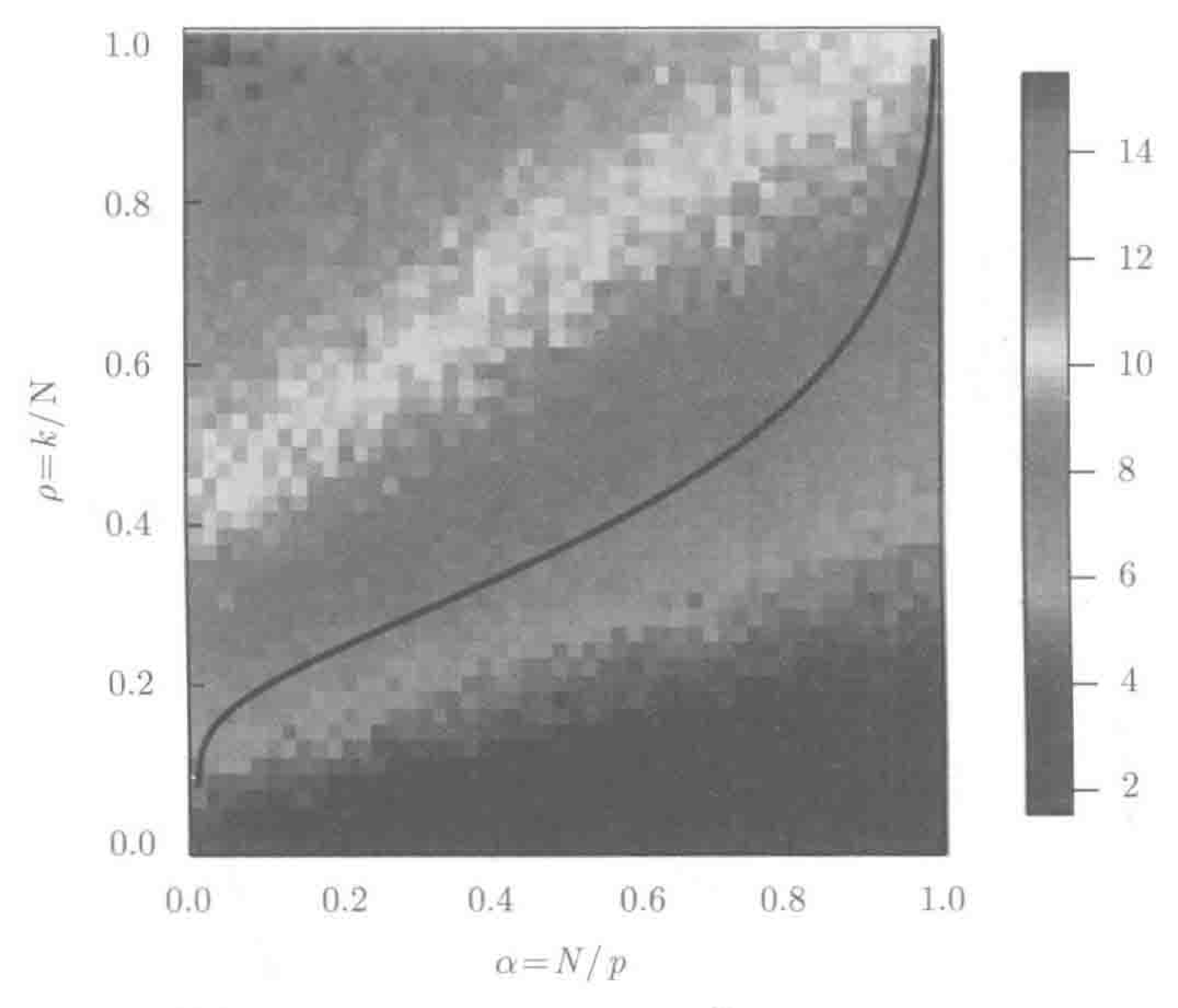



图 11-4　仿真实验：10 个样本上的误差 $\|\widehat{\beta}-\beta^*\|_2$ 中值，有界（11.18）


站在巨人的肩上
Standing on Shoulders of Giants
TURING
图灵教育
iTuring.cn


站在巨人的肩上
Standing on Shoulders of Giants
TURING
图灵教育
iTuring.cn